IDENTIFYING, INVESTIGATING, AND CLASSIFYING DATA ERRORS: AN ANALYSIS OF CLINICAL RESEARCH DATA QUALITY FROM AN OBSERVATIONAL HIV RESEARCH NETWORK IN LATIN AMERICA AND THE CARIBBEAN

By

Stephany Duda

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in

Biomedical Informatics

May, 2011

Nashville, Tennessee

Approved:

Professor Cynthia S. Gadd

Professor Daniel R. Masys

Professor S. Trent Rosenbloom

Professor Bryan E. Shepherd

Professor Sten H. Vermund

ACKNOWLEDGEMENTS

I would like to thank the National Institutes of Health and the National Institute of Allergy and Infectious Disease, whose CCASAnet grant (#U01 AI069923) has made this research possible.

I'd like to extend particular thanks to the five members of my Ph.D. committee, who have been invaluable in the preparation of this work. Dr. Daniel Masys, the CCASAnet Primary Investigator, initiated the program of data quality audits from which this body of work emerged. He generously mentored me on data quality, international collaborations, and life for all five years of the grant. Dr. Bryan Shepherd patiently explained statistical methods and challenged me to find new applications for our audit findings. I hope I can learn to emulate his way of asking even the most challenging questions kindly and making everyone feel valued for their project contributions. Dr. Trent Rosenbloom offered an essential outside perspective to this work and encouraged me to consider research implications beyond global health. I will always appreciate his thoughtful comments and sense of humor. Dr. Sten Vermund made time for my projects in the midst of his demanding schedule, offered constructive advice, and inspired me with his confidence in both my abilities and the value of this research. I am grateful he gave me an opportunity to write the informatics component of a grant and mentored me on the process. Dr. Cindy Gadd, my dissertation advisor and committee chair, supported me unfailingly through all stages of this dissertation. Her depth of knowledge, insight, and patience were reflected in every step of the project. She spent long meetings discussing work with me, long evenings proofreading, and kept me focused on a schedule for success. But in the end I found she cared more about me as a person than about the timeline, and it meant the world to me. Truly, it has been an honor to work with all of you.

I would like to thank all the participants in CCASAnet, especially our director of clinical research, Dr. Cathy McGowan. I could not have asked for a better mentor, travel companion, and friend with

whom to share this work. I am also grateful to Melanie Bacon, the NIAID Project Scientist for CCASAnet, for her guidance on the presentation of our audit results, and to the investigators at each of the CCASAnet sites for their willing participation.

I would like to thank all the members of the Department of Biomedical Informatics for their advice and assistance. Ms. Rischelle Jenkins was an unending source of reassurance, motivation, and cheerfulness. Thank you also to everyone who helped me through the last month of preparation: Judith, Kim, Laurie, Kat, Naqi, Josh, Firas, Maggie, Laura, Victor, Betsy, Ray, Alli, Jeff, and Barbara. I am so lucky to have such supportive and caring friends.

Finally, I would like to thank my parents, Hilary and Francis, and my brother Brendan. Their care, support, generosity, strength, and encouragement during the hardest times have seen me through this PhD. I wouldn't have been able to finish this without you.

Thank you, a hundred times over.

TABLE OF CONTENTS

Page

ACKNOWLEDGEMENTS ii

TABLE OF CONTENTS iv

LIST OF TABLES vii

LIST OF FIGURES viii

LIST OF ABBREVIATIONS ix

Chapter

I. INTRODUCTION 1

Dissertation Contents 3

II. BACKGROUND 4

Definitions of Quality 4
Impact of Errors 5
Quality of Medical Data 6
Quality Control in Medical Research 8
Data Auditing 10
Barriers to Audit 12
Study Setting: CCASAnet 13
CCASAnet Audit Process 16

III. MEASURING THE QUALITY OF OBSERVATIONAL DATA IN AN HIV RESEARCH NETWORK 18

Introduction 18
Methods 19
Study Setting 19
Data Audit Preparation and Process 21
Error Classification 22
Statistical Analysis 22
Results 23
Record Availability 23
Error Rates 23
Re-Auditing of Sites 28
Discussion 29
Summary of Findings 29
Recommendations 30

Limitations 32
Conclusion 32

IV. INVESTIGATING PERCEIVED REASONS FOR DATA QUALITY VARIATIONS 34

Introduction 34
Methods 35
Subjects and Settings 35
Survey Design 36
Data Analysis 37
Results 37
General Questions 38
Weight Measurements and Dates 39
Laboratory Results and Dates 40
Antiretroviral Regimens and Dates 41
Discussion 43

V. DEVELOPING A COMPUTER-ASSISTED TOOL FOR SOURCE VERIFICATION DATA AUDITS 45

Introduction 45
Motivation 48
Study Setting 48
Paper-based Audit Process 48
Limitations of Paper-based Audit 49
Key Attributes of an Audit Tool 53
Networking 53
Audit Data Management 54
Standardized Assessment of Errors 54
Audit Decision Support 54
Results Reporting 55
System Design 55
Setting 55
Implementation 57
Application Walkthrough 58
User Experience 66
Discussion 67
Summary of Findings 67
Limitations 68
Conclusion 69

VI. EVALUATING DATA QUALITY DIMENSIONS IN THE CONTEXT OF CLINICAL RESEARCH DATA AUDITS 71

Introduction 71
Study Subject: CCASAnet Audit Datasets 72
Methods 74
Results 75
Discussion 80

Summary of Findings 80
Limitations 81
Conclusion 82

VII. DISCUSSION 83

Summary of Findings 83
Study Limitations 84
Study Implications and Future Work 86

Appendix

A. EXAMPLE OF AUDIT FORM AND ERROR CODING 87
B. EXAMPLES OF CAAT METADATA AND DATA SPECIFICATIONS 89

REFERENCES 90

LIST OF TABLES

Table Page

1. Countries and clinics participating in the CCASAnet collaboration 15
2. Characteristics of data abstraction and management at audit sites A-G 20
3. Availability of randomly selected clinical records requested by the audit team according to site 23
4. Total number of audited variables and percentage of erroneous data by variable type during initial CCASAnet audits at seven sites. 24
5. Variable counts and error rates by data category during initial and follow-up audits at Site B 29
6. Questions included in data quality surveys administered to sites A-G 37
7. Groups involved in data abstraction, as reported by sites A-E, G 38
8. Groups involved in research data entry, as reported by sites A-E, G 39
9. Desiderata for a computer-assisted tool for data audits 56
10. EORTC audit codes 74
11. Sources and types of audit documentation included in data quality review 75
12. Types of data quality observations not captured by the standard audit protocol, their frequency, and related data quality dimensions 77
13. Example of the Demographics and Laboratory Data sections of a completed audit form 87
14. Example of the Antiretroviral Regimens sections of a completed audit form 88
15. Example of CAAT metadata specifications 89
16. Example of CAAT data specification 89

LIST OF FIGURES

Figure Page

1. Comparison of data collection in clinical trials and research networks 10
2. Member countries of the Caribbean, Central, and South America network for HIV Epidemiology .. 14
3. Breakdown of error rates by error type for birthdate, gender, weight, and weight date variables from audit sites A-G. 25
4. Breakdown of error rates by error type for CD4 and viral load-related values from audit sites A-G. 26
5. Composition of error rates for antiretroviral regimen data from audit sites A-G. 27
6. Composition of error rates for antiretroviral regimen start and stop dates from audit sites A-G..... 28
7. A neatly completed paper form from a CCASAnet site audit 49
8. A messy paper form from a CCASAnet site audit 51
9. Screenshot of the homepage of SimpleAudit, a prototype computer-assisted data auditing tool ... 59
10. Screenshot of the data import screen 59
11. Screenshot of the audit record selection page 61
12. Screenshot of the listing of records to be audited 61
13. Screenshot of the data audit interface 62
14. Screenshot of the record completion screen 63
15. Screenshot of the audit summary report 64
16. Automatically generated audit forms 65

LIST OF ABBREVIATIONS

Abbreviation	Definition
AIDS	Acquired Immune Deficiency Syndrome
ARV	antiretroviral medication
CAAT	computer-assisted audit tool
CCASAnet	Caribbean, Central, and South America network for HIV epidemiology
CR	clinical record
CD4	$CD4^+$ lymphocyte count
CSV	comma separated values
DB	database
DCC	data coordinating center
DQA	Data Quality Assessment
EMR	electronic medical record
EORTC	European Organization for the Research and Treatment of Cancer
GCP	Good Clinical Practice
GDMP	Good Data Management Practice
HAART	highly active antiretroviral therapy
HIV	human immunodeficiency virus
ID	identification number
IeDEA	International epidemiologic Databases to Evaluate AIDS
IQR	interquartile range
ISO	International Organization for Standardization
NIAID	U.S. National Institute of Allergy and Infectious Diseases
NICHD	U.S. National Institute of Child Health and Human Development

NIH ..... U.S. National Institutes of Health

PEPFAR .....The United States President's Emergency Plan for AIDS Relief

RDQA..... Routine Data Quality Assessment

SOP .....Standard Operating Procedure

VL .....HIV viral load

CHAPTER I

INTRODUCTION

High quality data are essential to the accuracy and validity of clinical study results. Assuring data quality has been a particular emphasis of clinical trials, where extensive personnel training and data monitoring programs are built into the study protocol in an effort to prevent scientific misconduct and ensure compliance with the International Conference on Harmonization's guidelines for Good Clinical Practice [1]. Yet clinical trials can be elaborate and expensive and the cohorts are not always large or varied enough to answer broad research questions. As a result, researchers and funding agencies seek to leverage existing clinical care data by pooling datasets from multiple sites [2]. Indeed, the United States of America's National Institutes of Health (NIH) have indicated an interest in promoting and expanding such clinical research networks by featuring them as a cornerstone of the NIH Roadmap for Medical Research [3,4]. The U.S. Nationwide Health Information Network, a standards initiative for health information exchange over the Internet, supports complementary standards for both clinical care and clinical research data in order to encourage and support the reuse of healthcare data for observational studies and population monitoring [5].

Medical research is experiencing a simultaneous upsurge in international research collaborations [6,7]. Membership in multi-national research networks has grown exponentially and publications by multi-national research teams receive more citations than similar work from domestic collaborations [8,9]. These trends combine in the increased reuse of clinical care data for international research collaborations. Data collected during routine patient care are readily available and relatively

inexpensive to acquire [10], so even clinical sites in resource-limited settings are contributing data to shared repositories or multi-site datasets [11,12].

Unfortunately scientists seldom investigate the quality of such "secondary use" data as thoroughly as data generated in clinical trials or similarly regulated studies [13]. Some research groups rely on data cleaning performed at the data coordinating center to detect data discrepancies, or request that their participating sites perform regular quality self-assessments [14,15]. Given time and funding restrictions and a dearth of data management personnel in academic centers, it is likely many groups simply accept secondary use data as is.

We believe significant challenges to high quality data exist within such international, multi-site research networks, but that these issues can be remedied through well-planned, cost-effective quality control activities. Our projects investigate the necessity of data quality assessments for observational networks, as well as means of identifying and evaluating data errors and improving the audit process. This work encompasses four aims:

Aim 1: Evaluate through a series of source verification data audits the accuracy and completeness of observational data submitted to an international, multicenter HIV research network;

Aim 2: Investigate reasons for data quality variations, as perceived by the local research teams;

Aim 3: Develop a computer-based audit tool for source verification data audits; and

Aim 4: Evaluate data quality dimensions in the context of clinical research data audits.

The findings demonstrate that data quality control activities should not be limited to clinical trials and that investigators who reuse clinical care data must be particularly vigilant regarding its

quality. The causes of data errors can vary greatly, and collaborative research networks – especially those in international settings – should consider implementing an audit program to evaluate the reliability of different data sources and correct discrepancies in data that have already been collected. Such audits can be made more efficient by using computerized tools and flexible error metrics.

Dissertation Contents

This report comprises six chapters describing four associated research aims. This chapter (Chapter I) introduces the research problem. Chapter II defines the concept of data quality, provides an overview of quality control and data auditing in medical research, and introduces the international, multi-site, HIV research network that serves as the setting for the subsequent data quality-related projects. The network's data auditing process and the findings from the first seven site audits are discussed in Chapter III, along with the audit team's recommendations for data quality improvement. Chapter IV presents the varied reasons for data errors at each network site, as elicited from clinicians and data entry personnel by an anonymous survey. We explore the feasibility of improving such audits using computer-assisted auditing tools in Chapter V. Chapter VI discusses extant error taxonomies used in data auditing and the types of information they failed to capture during our on-site audits, while identifying additional relevant dimensions of data quality. The concluding chapter (Chapter VII) summarizes the research findings and discusses additional directions for this work.

CHAPTER II

BACKGROUND

Definitions of Quality

The value of conclusions drawn from data relies heavily, though not exclusively, on the data's quality. High quality data are crucial to sound decisions in diverse domains, including financial transactions, political and military policy, product manufacturing, patient care, and biomedical research. Yet scientists and government regulators have hitherto not adopted a universal definition of data quality. The U.S. Department of Health and Human Services' "Guidelines for Ensuring and Maximizing the Quality, Objectivity, Utility, and Integrity of Information" – a product of the U.S. Data Quality Act – proposed that high quality data were useful to their intended audience, presented in context and in a complete and unbiased manner, well-sourced and documented, and secured from tampering or unauthorized changes [16]. An alternate definition emerged from the U.S. Institute of Medicine's 1999 Roundtable on Research and Development of Drugs, Biologics, and Medical Devices, which stated "'high-quality data' refers to data that can be used without further revisions or data that will produce conclusions and interpretations that are equivalent to those that would be derived from error-free data, that is, data that are accurate, reliable, and fit for use" [17]. Although the International Organization for Standardization (ISO) has begun development of a standard for data quality, ISO 8000, applications of its as-of-yet-unfinished specifications have received no attention in the scientific literature [18,19].

The common characteristic among these and other published data quality definitions is that data quality is a compound concept. Over twenty-six distinct aspects of quality are described in the literature, occasionally with identical names and conflicting definitions [20]. The most frequently

reported “dimensions of data quality” include *accuracy* (the extent to which a recorded data point conforms to the true value), *completeness* (the extent to which necessary data that could have been recorded have been recorded), *timeliness* (the rapidity with which data are entered into the database after they are generated), *conciseness* (the degree to which data have been represented in a compact format with minimal redundancy), *objectivity* (the extent to which data are free from bias), *transparency* (the detail with which the data collection process is described in accompanying documentation, supporting information reproducibility), *reputation* (the degree of trust in the data’s origin), *security* (the extent to which the data have been protected from accidental or malicious manipulation), and *relevance* (the extent to which the data fit the researcher’s needs)[20-25].

Not every application of data requires high quality measures in each of these dimensions. For example, district-level data are unnecessarily precise for a graphic of population size by country, and timeliness is an unimportant data quality dimension for a retrospective study of cancer incidence two decades ago. In assessments of clinical research data, accuracy and completeness of the dataset are considered the most critical markers of data quality. When a dataset registers high quality in all dimensions relevant to the researcher, it is deemed “fit for use.” This viewpoint is summarized in Arthur Chapman’s *Principles of Data Quality*: “in a database, the data have no actual quality or value; they only have potential value that is realized when someone uses the data to do something useful”[26].

Impact of Errors

Despite disagreements on definitions of high quality data, researchers agree poor quality data is costly. When investigators study datasets that are not fit for use, the data’s deficiencies may foil analyses, lead to biased conclusions, and possibly trigger expensive quality interventions. In application, this misinformation yields manufactured products that fail to meet specifications, ill-tailored business

services, and harmful or suboptimal health care. Poor customer data quality, for example, has a negative impact on customer-business relationships [27] and a survey by Information Impact International, a U.S. consulting firm, indicated low quality data may cost organizations 10% of revenues [28]. In economics, data errors have led to misclassification of 34% of countries on the Human Development Index [29]. Applications of flawed data can endanger patient health: missing data from electronic patient records have caused a clinical diagnostic support system to generate "inappropriate and unsafe recommendations" on the risk of gastrointestinal bleeding in 77% of patient encounters [30]. Implemented findings from chemotherapy studies using falsified data have resulted in serious patient harm [31].

Data errors also impact biomedical research. A study by Ancukiewicz et al. assessed the effect of medical record abstraction and data entry errors on an analysis of radiation oncology data. An expert validation of all variables revealed that 2.7% of data elements contained a data abstraction or entry error and that these errors resulted in statistically and clinically significant differences in the survival rates of the analyzed cohort [32]. Data entry errors alone, particularly errors in binary variables, can corrupt the results of epidemiologic studies [33]. And even moderate error rates of 1-5% in phenotype/genotype datasets can obscure the results of genetic linkage and association studies [34,35].

Quality of Medical Data

Despite the consequences of error, medical data collections suffer from high rates of incomplete and incorrect data and concerns about the quality of medical data are reflected in studies of disease registries, electronic medical records, and clinical trials datasets. Cancer registries in Norway, Scotland, have reported 1.2% of requested fields are missing and 2.8% contain clinically significant errors [36,37]. Extensive reviews of cancer research datasets revealed that 1-2.8% of variables did not match data in

the clinical record [38,39]. A survey of antiretroviral therapy (ART) programs in low-income countries found on average 10.9% of key patient care variables were missing from the local electronic medical database [40]. Similarly, an evaluation of 29 medical record elements available in a Dutch university hospital information system found only 19 were usable without manual reformatting or data cleaning and that most errors stemmed from inaccurately recorded times of treatment events, missing symptom or diagnosis data, incorrect evaluations of illness severity, and prescription data [41].

As illustrated by the Dutch study, some data types may be more prone to error than others. Demographic and medication data tend to contain fewer errors than patient problem lists or diagnoses, according to a landmark review of data accuracy in electronic medical records [42]. Not all patient intake data receives the same attention as demographics, however; lifestyle data such as history of smoking and alcohol use, socioeconomic status, patient occupation, and ethnicity are frequently overlooked [43]. Within data forms, answer options associated with more complex patient treatments are more prone to error, as are data that require interpretation or physician assessment [44].

Errors can be introduced during multiple stages of the data collection process. The events that cause data distortion or loss include *data capture* (when data are generated and recorded), *record abstraction* (when data are transcribed from the medical record to a case report form, or CRF), and *data entry* (when data are copied from the form into the study database.) Of these, data abstraction is considered the primary source of error, resulting in ten times as many errors as CRF-to-database entry [45]. Concerns about medical record abstraction solidified after a 1974 Institute of Medicine report that concluded “diagnosis-specific discrepancies [produced by record abstraction] are of sufficient magnitude to preclude use of such data for detailed research and evaluation” [46]. More recent studies of data abstraction and entry error rates reported 11.5% to 31% of all requested data fields were overlooked, improperly interpreted, or incorrectly transcribed [47,48].

Pharmaceutical and medical device companies seeking U.S. government approval for their products are held to higher data quality standards, just one to five errors in 1,000 data fields is considered an acceptable error rate [49].

Quality Control in Medical Research

The International Conference on Harmonisation (ICH) publishes a standard for ethical and scientific quality in clinical trials that involve human participants. These standards, called Good Clinical Practice (GCP) principles, were developed in response to exposés of study fraud and pharmaceutical trials that led to patient harm [50]. GCP outlines requirements for designing, conducting, and reporting the results of clinical trials [1]. For investigators following GCP, quality assurance and control activities help to ensure scientific validity, monitor patient safety during clinical trials, catch systematic data collection problems and prevent future errors, increase public confidence in the study conclusions, and meet requirements for regulatory approval [51]. Studies often conduct duplicate data entry to reduce the frequency of typographical errors in the final dataset [52], and may integrate electronic data capture software to eliminate paper case report forms. Such software usually includes range and valid value checks for data entry as well as automated database checks for internal consistency [53].

Extensive and regular training of study investigators and staff serves to further reduce errors in data collection [54]. Protocols may specify strict adherence to published standard operating procedures for patient interaction and data collection and storage. Many multicenter trials support a phone hotline for questions from sites and maintain frequent email contact to sustain motivation and attention to detail among study personnel [53].

Source documents are the gold standard for recorded information in a clinical trial [55]. The ICH guideline defines them as “original documents, data, and records” from which the study data derive,

including hospital records, clinical and administrative office charts, laboratory reports, pharmacy dispensing records, recorded data from automated instruments, and records from outside clinics, as well as certified copies or transcriptions of such documents [1]. Missing source documents and frequent data errors amount to major violations of clinical trials data quality codes [56].

As medical record abstraction frequently generates errors, clinical trials that rely on abstraction data sometimes implement additional quality interventions, including individual data abstractor training sessions, a thorough training manual that accompanies the Standard Operating Procedures for data collection, standardized record abstraction examples, on-site visits, occasional double-abstraction sessions for pairs of auditors, weekly team conference calls, and a rapid cycle of data submissions from sites followed by data cleaning requests from the coordinating center [57]. These interventions effectively reduce data error originating from medical record abstraction, but the additional time and funding they require ensures that only the most rigorous studies implement such measures.

Observational research networks face quality challenges that arise from repurposing data collected primarily for clinical care. In protocol-driven prospective studies, such as clinical trials, methods of error detection include prospective quality assurance activities that take place before and during a trial, including case report form design and testing, standardization of data collection tools and procedures, and ongoing training of study personnel [58]. Studies using pre-existing data are limited to quality control activities that occur once active data collection in a study has ceased, as illustrated in Figure 1.

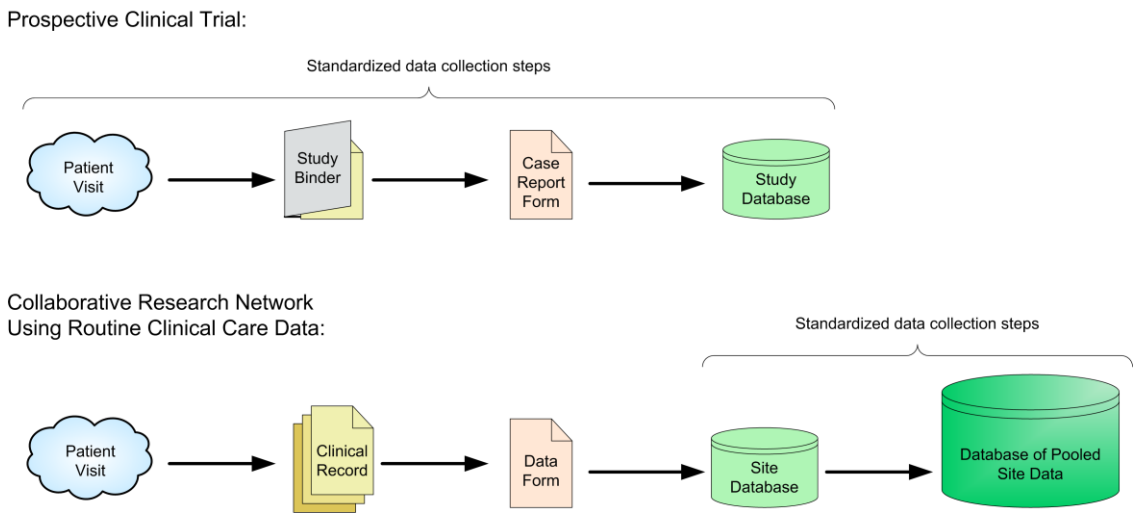


Figure 1: Comparison of paper-based data collection in clinical trials and research networks
This figure illustrates the process of paper-based data collection in prospective clinical trials and in collaborative research networks reusing clinical care data. In a clinical trial, all steps of the data generation process are supervised by the study coordinators. In a collaborative research network, only the final step of dataset creation – the pooling of multiple site datasets at the data coordinating center – is subjected to a standardized process.

Problems with conflicting data collection methodologies and data quality standards are amplified in international and low-resource settings, where different resources and cultural norms affect the conduct of research and delivery of healthcare [59]. The more sites differ in their data collection procedures and data storage formats, the harder it becomes to merge the data and derive accurate study results [60].

Data Auditing

Auditing is an established technique for evaluating and improving the quality of products, services, or information, and has been a staple of quality control activities for over a century [61,62]. Audits take many different forms depending on the domain: in accounting, they identify fraud; in manufacturing, audits help assess both the quality of a product lot and the producer's compliance with Good Manufacturing Practice; and in information security, audits allow for the

inspection of the security and reliability of computer systems and of the information they contain [63,64].

In medicine, researchers have employed audit techniques to detect inconsistencies in terminologies, evaluate the quality of patient care and verify that medical services are properly documented, coded, and billed [65-68]. Audits of health care outcomes in Africa provide insight into regional policies on HIV/AIDS care [69]. The U.S. Federal Drug Administration also requires auditing of many clinical trials to ensure that the operators of the trial are properly monitoring patient safety, accurately recording data generated by the study, and adhering to the study's protocol and Good Clinical Practice [1,70]. The U.S. Veterans Administration conducts manual quality assessments of computerized patient records every three months [71,72], while research groups like EuroSida and the European Organization for Research and Treatment of Cancer conduct yearly quality monitoring visits at centers contributing data to their cohorts [60,73].

Under ideal circumstances, clinical data audits are conducted by a team of trained auditors following a structured audit methodology for GCP, defined as a "systematic approach to auditing characterized by a prescribed, logical sequence of procedures, decisions, and documentation steps, and by a comprehensive and integrated set of audit policies and tools designed to assist the auditor in conducting the audit"[74]. This group is responsible for determining the goals of the audit, alerting auditees, and selecting audit records.

Once on-site, the audit team conducts a facility tour, interviews the study staff, and inspects the study regulatory binder. The majority of the audit is spent reviewing source documents such as the clinical record, laboratory reports, and pharmacy records for at least 10% of cases [53]. Auditors evaluate whether patient safety has been maintained during the trial and whether study data were

collected according to GCP guidelines. During the last day of the on-site visit, auditors complete an audit checklist and hold an exit interview with local clinical investigators to present the initial audit findings.

Barriers to Audit

Data audits are not cheap. In 2002, a single-site FDA audit for pre-marketing drug approval cost on average USD $7,350 for domestic locations, or $9,400 for audits at international sites, and required close to 100 auditor hours for preparation, inspection, and report writing [17]. In a large clinical trial involving 10,000 patients, 100% audit of all data points could incur costs in excess of two million dollars [53]. Resolving errors identified by data review consumes additional time; revising inaccuracies (for 0.44% of data elements) in 499 records from a head injury database took approximately 15 person hours, while filling in missing data (7%) took over 60 person hours [14].

However, science is not inherently self-correcting and efficient audits can be a cost effective means of identifying flaws in research [75]. Yearly audits of 10-20% of records is a cheaper solution than complete audits and still provides a reliably data quality benchmark. Clinical trials investigators have also found the cost of audits decreases over time as quality assurance procedures are reinforced during the conduct of a trial [53].

Local clinicians and data management staff often fail to notice the shortcomings of their data collection techniques; external observers can offer feedback that improves the correctness and completeness of future data collections [76]. Visiting individual sites in a multi-center research consortium also cultivates collegial working relationships among researchers and allows auditors to identify organizational and workflow practices that may influence the quality of data collection [60]. These insights can affect how individual site datasets are handled during data cleaning and analysis.

Academic research studies, however, may involve fewer data quality assurance activities than regulated clinical trials conducted by clinical research organizations, possibly because they lacks a supervising agency, a standard hierarchy of responsibility and command among investigators, and a client to assume the costs of quality assurance [77,78]. Researchers also resist conducting audits because they generate extra paperwork, consume funds that could otherwise go to new research, and are perceived as being excessive for unregulated studies while implying distrust of researchers' commitment, ability, and ethics [77]. Lack of time and motivation among investigators, unclear audit procedures, and difficulties with staff coordination frequently factor in the failure of voluntary audit programs [79].

Proposals to enforce routine auditing in academia include conducting independent audits of the scientific literature to ensure the validity of study results and the integrity of fellow investigators [80], requiring universities to conduct routine audits of data collections in research departments [81], and establishing research certification institutions that conduct random audits of 1% of NIH-sponsored studies [78]. There has been little discussion of who would fund such initiatives, however. Academic grants rarely include enough funds to establish an audit program [77].

Study Setting: CCASAnet

All four components of this research address the implementation and effectiveness of data auditing in an international, multicenter observational research network for HIV. The project began in mid-2006 when the International Epidemiologic Databases to Evaluate AIDS (IeDEA) initiative, funded by the U.S. National Institute of Allergy and Infectious Diseases (NIAID) and the National Institute of Child Health and Human Development (NICHD), established seven regional research networks for the collection and harmonization of global data on the epidemiology and treatment of HIV. Four of the

sponsored networks were located in Africa (East, West, South, and Central), one in Asia, one in North America, and one in Latin American and the Caribbean. In addition, the initiative includes the establishment of an international research consortium. Through these data centers and consortium, researchers will address unique and evolving research questions in HIV/AIDS that individual cohorts are unable to address.

The Caribbean, Central, and South America Network for HIV Epidemiology (CCASAnet) is one of the seven IeDEA regional collaborations and is the setting for all research reported here. The CCASAnet partnership combines the clinical and research expertise of investigators at seven HIV clinics in Argentina, Brazil, Chile, Haiti, Honduras, Mexico, and Peru with bioinformatics and biostatistics resources from Vanderbilt University in Nashville, Tennessee, USA. The seven member countries are labeled on the map in Figure 2.

Figure 2: Member countries of the Caribbean, Central, and South America network for HIV Epidemiology
Countries participating in the CCASAnet consortium include Argentina, Brazil, Chile, Haiti, Honduras, Mexico, and Peru.

Most CCASAnet sites are large, urban HIV clinics with associated academic or public hospitals that employ paper records for patient care. Clinic personnel transfer data from paper records into locally developed electronic databases for patient tracking and research. The names and locations of participating sites and the number of records in their research database are listed in Table 1.

Table 1: Countries and clinics participating in the CCASAnet collaboration
Sites in seven countries participate in the CCASAnet consortium. The numbers of records in each site's dataset are current as of November 2010.

CCASAnet Site	City	Participating Clinics	Records
Argentina	Buenos Aires	Fundación Huésped, Hospital Fernández, Centro Médico Huésped	8400
Brazil	Rio de Janeiro	Hospital Universitario Clementino Fraga Filho	600
Chile	Santiago	Fundación Arriaran	2800
Haiti	Port-au-Prince	GHESKIO	3700
Honduras	Tegucigalpa	Instituto Hondureño de Seguridad Social, Hospital Escuela	900
Mexico	Mexico City	Instituto Nacional de Ciencias Médicas y Nutrición Salvador Zubiran	400
Peru	Lima	Instituto de Medicina Tropical Alexander von Humboldt	3400

The CCASAnet collaboration encompasses six core objectives, to (1) develop a network of clinical sites in Latin America and the Caribbean, (2) assemble datasets from pooled routine care data from participating sites, for the purposes of answering questions that cannot be answered using datasets , from any single locale, (3) develop new biostatistical methods to handle such data, (4) conduct joint research projects that further the understanding of HIV and related diseases in the region, (5) work with sites to enhance their clinical research capacity, and (6) collaborate with other IeDEA regions on global projects in HIV research [82]. CCASAnet's clinical studies to-date have explored mortality during the first year of antiretroviral therapy among HIV-infected patients in the region, reasons for therapy changes, cancer incidence in the cohort, reasons for late initiation of antiretroviral therapy, and recurrence of tuberculosis among HIV+ persons [83-86].

CCASAnet Audit Process

Poor quality datasets could present significant difficulties for CCASAnet studies by preventing the identification of accurate subsets of patients for retrospective analyses, hampering the process of merging data from multiple sites, producing invalid conclusions from the research data, and making it unfeasible to compare results to findings in other settings. To identify problems with the reuse of routine patient care data, CCASAnet has instituted a process of periodic data monitoring visits at its member sites, using audit principles adapted from GCP guidelines. The CCASAnet Data Coordinating Center hosted at Vanderbilt University conducts project-driven audits when new datasets are submitted for proposed region-wide studies. Such audits help CCASAnet participants to identify sources of error in data collection, abstraction, and representation, and help the Data Coordinating Center (DCC) to determine the structure, quality, and reliability of the submitted data. The DCC also conducts a second round of supportive quality improvement-focused audits at sites where significant challenges to data capture and quality have been identified.

The audit cycle begins when a CCASAnet site submits data to the coordinating center for a region-wide project. A team from the DCC visits each of the CCASAnet participating sites and compares the contents of the electronic database the site has submitted to the DCC to local source documents. The source documents available at CCASAnet sites are mainly paper clinical records, though some sites maintain electronic laboratory, pharmacy, and patient medical record systems. On the final audit day, the audit team meets with the site PI and describes the preliminary findings of the data monitoring visit. The group discusses the site's data quality and the strengths and weaknesses of the current data collection approach. The audit team then presents an initial set of recommendations and discusses their feasibility with the local PI and data personnel.

After the audit team returns to the DCC, team members review the audit records, classify data errors, omissions, and inconsistencies, and tabulate the results by data category. They also compile comparison charts for individual records that present side-by-side the database and source document content, so local data personnel can see specific discrepancies. The audit team then composes a data audit report that contains the error tables and comparison charts and describes in detail audit findings and the DCC's recommendations. The DCC consults with site data personnel to adapt and implement the recommendations detailed in the audit report and solicits feedback from the sites about the data audit process. CCASAnet data audit preparations and processes are described in greater detail in Chapters III and V.

In summary, routine clinical care data, particularly data generated in resource-limited settings, is often of poor quality. Although such data can compromise clinical research findings, researchers rarely conduct extensive quality assessments, particularly when managing unregulated, observational studies. Auditing is an expensive but cost-effective method of determining the quality of research data.

CHAPTER III

MEASURING THE QUALITY OF OBSERVATIONAL DATA IN AN HIV RESEARCH NETWORK

Introduction

Accurate and valid HIV research results depend on high-quality clinical and laboratory data. Excellent patient care itself depends on accurate recording and transcription of such information [87]. Interventional clinical trials, such as those generating data for pre-marketing approval by regulatory agencies or those conducted by the AIDS Clinical Trials Group [88], follow careful data quality assurance procedures. This ensures that investigators comply with the International Conference on Harmonization's guidelines for Good Clinical Practice and that collected data reflect true measurements [1,89]. A data audit, in which an external review team compares a research dataset to the original data collection documents, is the standard method of assessing the quality of data in clinical trials [38,59,90,91].

Depending on the nature of the research question, researchers and funding agencies may use routine patient care data as a readily available and inexpensive supplement or alternative to data generated through prospective clinical trials [10,92]. Databases that pool such observational data from multiple, international sites have become particularly important resources for HIV/AIDS research due to increased interest in measuring global trends in the epidemic and the side effects and long-term outcomes of antiretroviral (ARV) therapy [11,93-96]. Routine medical care data, however, do not undergo the same stringent quality controls commonly applied in clinical trials. International multi-center HIV networks that use routine patient care data may be at higher risk of having data quality

issues because monitoring for quality at geographically distant locations is difficult, time-consuming, and expensive.

Many networks that repurpose routine medical care data for research rely on data cleaning and cross-referencing performed at the data coordinating center [40]. These techniques can confirm the data's internal consistency and identify missing values but cannot determine data accuracy and authenticity. Comparing research data to their source documents through audits is therefore an essential additional step in verifying the data's accuracy [97]. Despite the importance of using high-quality data, no multi-center observational HIV cohort has published research indicating that they conducted frequent source-to-database data comparisons. As a result, the quality of data and the accuracy of results from many multi-site HIV cohorts can be uncertain.

We evaluated the accuracy and completeness, as assessed by on-site data audits, of routine patient care data submitted for research by seven sites participating in CCASAnet.

Methods

Study Setting

The audits took place at all seven CCASAnet sites in Argentina, Brazil, Chile, Haiti, Honduras, Mexico, and Peru. To preserve the anonymity of the participating clinics, we have labeled them randomly as sites A-G.

All seven CCASAnet sites used paper medical records as the primary means of storing patient information. The paper clinical records generally contained a structured patient intake form followed by handwritten visit notes. Nursing staff at sites F and G maintained detailed drug dispensing forms that

were kept with the patient chart and used to verify drug prescriptions and dates. The different data collection practices and resources at each audit site are detailed in Table 2.

Table 2: Characteristics of data abstraction and management at audit sites A-G
A plus sign ("+") indicates that the site employed such personnel or used such a form, system, or database. A minus sign ("–") indicates the site did not display the listed characteristic.

	Sites						
	A	**B**	**C**	**D**	**E**	**F**	**G**
Structured visit form	+	–	–	–	–	–	–
Drug dispensing form	–	–	–	–	–	+	+
Electronic laboratory system	–	+	–	–	–	+	–
Locally developed and maintained database	+	+	+	+	–	+	+
Data manager	+	+	+	+	–	–	+
Full-time data abstraction team	–	–	–	+	–	–	+
Internal data audits	–	–	–	–	–	+	–

At all sites, a team of clinicians, data entry personnel, or administrative staff abstracted information from the paper medical record and entered the results into an electronic research database from which data were extracted for CCASAnet studies. Sites B and F had access to electronic laboratory systems that exported test results directly into their research databases. The majority of these databases were designed and maintained by local staff and implemented in Microsoft Access. One site used a commercial data warehousing service in place of an on-site database server. Two of the seven sites employed experienced data managers (B, C); three of the remaining sites had data managers without formal training (A, D, G.) Only Site C operated an extensive data center. Site F was the only site that actively conducted internal quality reviews of their research data.

Institutional Review Board approval was obtained locally for each participating site and for the DCC. Local centers de-identified all data before transmitting it to the DCC.

Data Audit Preparation and Process

Between April 2007 and March 2008, a team from the DCC -- including at least one HIV clinician and one informaticist -- conducted on-location audits of the datasets received from CCASAnet member sites. Our audit techniques involved verifying data integrity using source documents and were adapted from those used in clinical trials to ensure Good Clinical Practice (GCP) compliance [70].

The DCC data manager selected approximately 30 records at random from the database submitted by each participating site. We sent research identification numbers (IDs) for 20 records to the site ten days before the data monitoring visit to allow local data personnel to retrieve the records in advance. We requested the remaining ten records from the site investigators on the first audit day.

Our initial audits lasted two to three days per site. Sites for which we recommended major quality interventions were re-audited during the current or subsequent audit cycle. During each audit, we compared the contents of the study database to the local source documents and noted discrepancies and inconsistencies in individual data elements. The available source documents included paper clinical records and, where available, electronic laboratory, pharmacy, and medical record reports. We reviewed as many source documents as the sites could locate during the visit and consulted local site personnel for clarifications as needed. Audit findings were recorded on a structured paper audit form and later entered into an Excel spreadsheet for analysis.

The audited variables included those most relevant to proposed consortium studies: patient demographics, HIV-related risk factors, weight measurements, $CD4^+$ lymphocyte (CD4) counts, plasma HIV-1 RNA levels (viral load), all ARV regimens, and all dates associated with each measurement. When an individual ARV regimen was recorded as “current” or “ongoing” in the database, we verified that the patient was still taking the specified drug combination at around time the site data were submitted to the DCC.

Error Classification

After each audit, we reviewed the completed paper audit forms and categorized audit results using standardized audit codes from the European Organization for the Research and Treatment of Cancer (EORTC) [38]. The audit data were labeled *correct* (code 1) if the database values submitted to the DCC matched the values in the paper clinical record or other on-site source documents. *Major errors* (code 3) represented discrepancies between the clinical record and database values that the physician on the audit team deemed clinically meaningful. Other discrepancies were labeled *minor errors* (code 2, e.g., weight values rounded to the nearest integer, dates in the database within six days of the documented date). *Missing/missed data* (code 4) included values for requested information (e.g., baseline weight) that auditors found in the clinical record but had been left blank in the database. Values that existed in the submitted database but not in the clinical record were labeled *sourceless* (code 5). One auditor performed the initial classification, a second auditor reviewed all classifications, and all disagreements were resolved by joint record review. A coded sample record is presented in Table 13 and Table 14 of Appendix A.

Statistical Analysis

We calculated error rates by dividing the number of erroneous clinical record/database value pairs (codes 3, 4, and 5) by the number of audited value pairs. We compared error rates across variables using a generalized linear mixed model to account for correlations between variables from the same record and within the same clinic. Analyses comparing major error rates (code 3) between variables did not include missing or sourceless errors (codes 4 and 5) in the denominator.

Results

Record Availability

We requested a total of 208 randomly selected patient records during seven data audits at Sites A-G. Of these 208 records, 16 could not be located or were unavailable because they were needed for patient care. The majority of missing records (11 of 16) were from Site A. We reviewed 184 of the remaining 192 charts (eight charts were not audited because of time constraints) comprising 4,223 unique data points. The number of unavailable, available, and audited charts by site is shown in Table 3.

Table 3: Availability of randomly selected clinical records requested by the audit team according to site

Audit Site	Total charts requested	Charts available and audited	Charts available but not audited	Charts unavailable
A	40	29	0	11
B	28	23	3	2
C	17	17	0	0
D	28	27	0	1
E	33	27	5	1
F	35	35	0	0
G	27	26	0	1
Total	**208**	**184**	**8**	**16**

Error Rates

The dataset of all audit results contained 3,581 correct data points, 66 minor errors, 171 major errors, 274 missing values, and 131 sourceless values. Minor errors – which were not counted towards error rates – included dates that were shifted by a few days (45%), inappropriately rounded weight measurements (36%), and weight, CD4, and viral load values where a probable typographical slip

resulted in small value differences (19%). Table 4 shows the number of variables audited at each site and their specific error rates.

Table 4: Total number of audited variables and percentage of erroneous data by variable type during initial CCASAnet audits at seven sites.
This table shows the number of variables audited in each of eleven categories of data, including gender, birth date, weight, CD4 count, viral load, antiretroviral (ARV) regimens, and all associated dates.

	Audit Sites[a]													
	A		**B**		**C**		**D**		**E**		**F**		**G**		**All**	
	N	%err	*N*	%err	*N*	%err	*N*	%err	*N*	%err	*N*	%err	*N*	%err	*N*	%err
Variables																
Gender	29	0%	23	0%	17	0%	27	0%	27	0%	35	0%	26	0%	184	0%
Birth date	29	7%	23	9%	17	0%	27	19%	27	0%	35	0%	26	0%	184	5%
Weight	29	31%	37	41%	55	11%	26	38%	27	93%	268	1%	45	2%	487	14%
Weight date	29	21%	37	30%	55	15%	26	38%	27	100%	268	0%	45	9%	487	14%
Laboratory data																
CD4	29	14%	33	21%	31	6%	96	13%	132	5%	134	1%	88	5%	543	7%
CD4 date	29	21%	33	27%	31	10%	96	16%	132	17%	134	1%	88	8%	543	12%
Viral load[b]	29	7%	26	42%	0	--	57	25%	120	7%	112	1%	84	4%	428	9%
Viral load date[b]	29	17%	26	42%	0	--	57	28%	119	13%	112	0%	84	7%	427	12%
Antiretroviral regimen data																
Regimen	46	11%	54	26%	23	13%	38	21%	49	22%	67	7%	47	19%	324	17%
Start date	46	28%	54	56%	23	13%	38	32%	49	39%	67	12%	47	26%	324	30%
Stop date	30	27%	54	50%	7	29%	38	29%	49	33%	67	10%	47	38%	292	30%
All	354	17%	400	34%	259	10%	526	21%	758	20%	1299	2%	627	10%	4223	14%

[a]Columns contain the counts for each site (*N*), along with the percentage of data that was labeled "in error" by auditors (%err). The reported percentage of erroneous data includes incorrect, missing, and sourceless values (error categories 3, 4, and 5), but not minor errors.
[b]Site C did not submit any viral load data.

All audited instances of patient gender were correct, and all birth dates were correct at four of seven sites. At the remaining three sites, 7-19% of the birth dates recorded in the site database differed by a week or more from values in the clinical record. Weight values and their associated dates were entered in the database with few or no errors at Sites F and G. Error rates at the remaining sites ranged from 11-93% for weight measurements and 15-100% for weight dates.

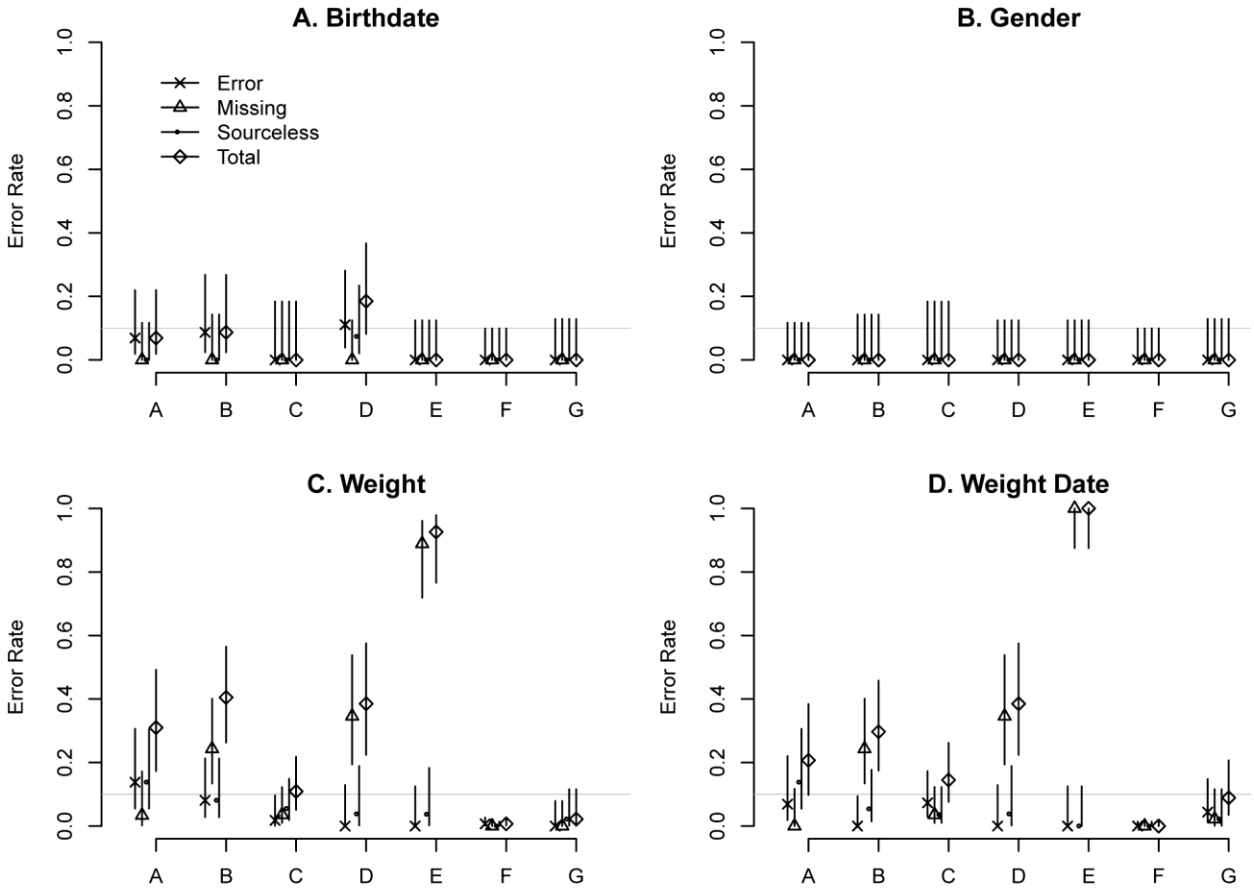


Figure 3: Breakdown of error rates by error type for birthdate, gender, weight, and weight date variables from audit sites A-G.
Values are shown with bars representing 95% confidence intervals. A gray line marks a 10% error rate.

At the three sites with the highest error rates, weight data missing from the database were the primary cause of error (24-89% missing weight measurements and 24-100% missing weight dates). The data error rates due to incorrect, missing, and sourceless data for birthdate, gender, and weight variables are depicted in Figure 3.

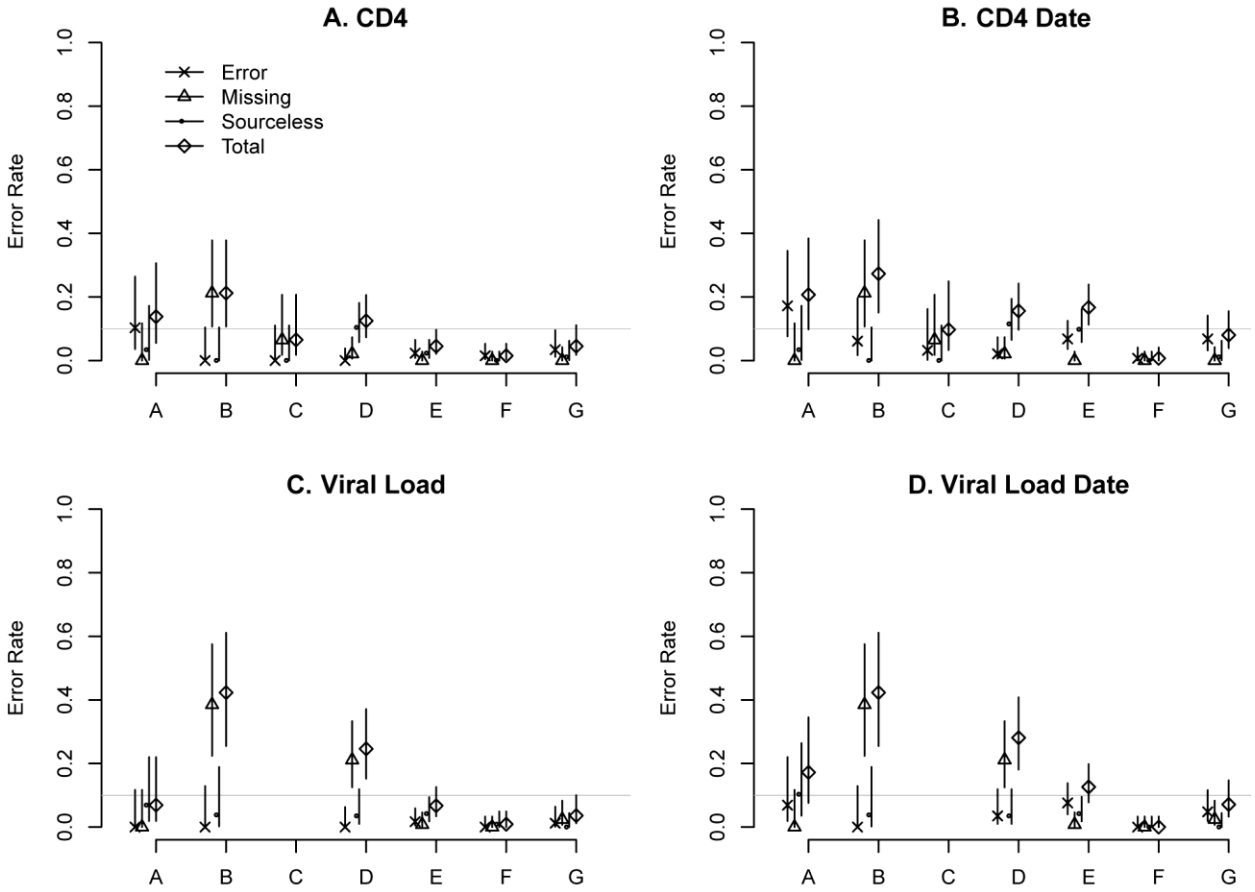


Figure 4: Breakdown of error rates by error type for CD4 and viral load-related values from audit sites A-G.
Values are shown with bars representing 95% confidence intervals. A 10% error rate is indicated with a gray line.

Sites varied less in error rates for laboratory values: CD4 count (1-21%), CD4 dates (1-27%), viral load (1-42%), and viral load dates (0-42%). For CD4-related values, the dominant error type varied by site. Site A's CD4-related errors primarily were due to incorrect information, Site B's errors to missing values, and errors at Sites E and F to data without source documents. For viral load measurements and dates, Sites B and D presented the highest error rates, primarily due to missing information. The composition of error rates is depicted in Figure 4. For the records with major errors (code 3) in laboratory data, the median and interquartile range (IQR) for absolute differences between the database and chart values were 20 cells/mm^3 (9-101 cells/mm^3) for CD4 count, 23 days (10-56 days) for

CD4 date, 65,200 copies/ml (47,000-145,600 copies/ml) for viral load, and 22 days (11.5-50 days) for viral load dates.

Errors in antiretroviral data according to audit site and error type (and their 95% confidence intervals) are shown in Figure 5. For 7-26% of ARV regimens, the drug combinations were missing from the site's submitted database, incorrectly entered into the database, or not substantiated by content in the clinical record. All sites had overlooked several old and intermittent drug regimens when abstracting patients' treatment histories for the datasets submitted to the DCC. Such missing data were least frequent at Sites D and F.

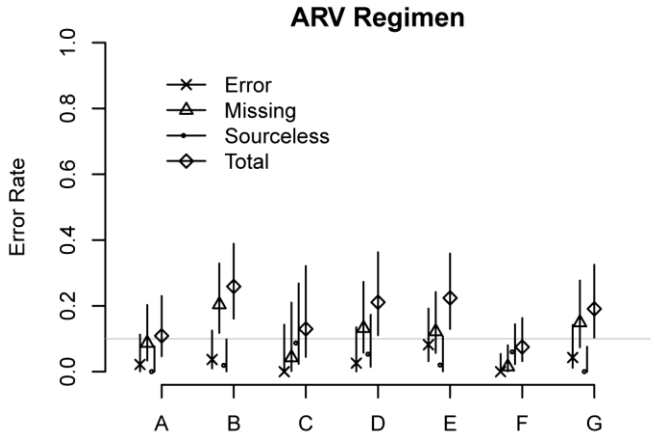


Figure 5: Composition of error rates for antiretroviral regimen data from audit sites A-G.
Values are shown with bars representing 95% confidence intervals. A 10% error rate is indicated with a gray line.

The start and stop dates of antiretroviral regimens appeared to have higher error rates than the regimens themselves, with rates ranging from 10-56%. Figure 6 depicts rates of erroneous ARV dates, which were in excess of 10% at most sites. There was no difference in error rates when comparing stopping and starting dates (P >0.25 for both major errors alone and overall errors.) For records with major errors (code 3), the median absolute difference between start/stop dates found in the chart and those recorded in the database was 88 days (IQR: 31 – 365 days).

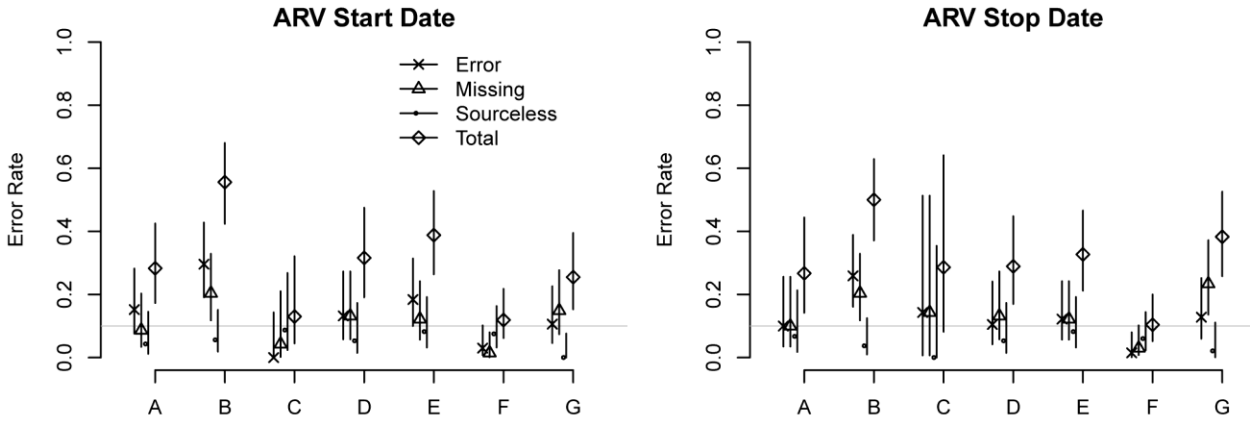


Figure 6: Composition of error rates for antiretroviral regimen start and stop dates from audit sites A-G.
Values are shown with bars representing 95% confidence intervals. A 10% error rate is indicated with a gray line.

Overall, the error rates for ARV regimen start and stop dates were higher than the error rates for all other non-ARV dates, including the dates of weight, CD4 count, and viral load measurements (*P* <0.001 for all.) For CD4 count, viral load, and ARV data, the associated dates had a higher rate of major errors (category 3) than the actual values (*P* <0.001 for all sites), whereas weight values had more major errors than weight dates (*P* = 0.028 for all sites).

Re-Auditing of Sites

After the findings of the initial audit, Site B cleaned and reabstracted their study data and quickly submitted an updated version, allowing us to reaudit the site at the end of the initial audit cycle. We reviewed 26 randomly selected records with 463 variables during this second audit at Site B; four additional records we requested were not available during the audit period. We observed reduced error rates in all variable categories. The overall error rate dropped from 34% to 17% (Table 5).

Table 5: Variable counts and error rates by data category during initial and follow-up audits at Site B

	Initial Site Audit		Follow-up Site Audit	
	N	%err	*N*	%err
Variables				
Gender	23	0%	26	0%
Birth date	23	9%	26	8%
Weight	37	41%	42	26%
Weight date	37	30%	42	21%
Laboratory data				
CD4	33	21%	35	6%
CD4 date	33	27%	35	6%
Viral load	26	42%	32	16%
Viral load date	26	42%	32	13%
Antiretroviral regimen data				
Regimen	54	26%	65	12%
Start date	54	56%	64	23%
Stop date	54	50%	64	33%
All	400	34%	463	17%

Discussion

Summary of Findings

Our data audits revealed substantial error rates in data submitted by all seven participating clinics. The majority of errors were due to measurements found in clinical records that were not entered into the database, laboratory values with no source documents, and incorrect antiretroviral regimens. Dates were especially prone to error, and sites had the most difficulty accurately capturing antiretroviral drug regimens and their associated dates. Most sites had error rates above 10% for ARV regimens and dates. These findings would trigger strict quality interventions in prospective clinical trials, which typically require fewer than 50 errors per 10,000 fields (<0.5% error rate) [49]. In the context of clinical trials, however, source-to-database audits like those described here generally report similarly high error rates when compared to case-report-form-to-database audits [45].

We found that data inconsistencies resulted from how the sites recorded information in the clinical record, how they abstracted data for research, and how they entered, stored, and formatted the data in the electronic database. Many laboratory dates could not be confirmed because the original laboratory reports had been discarded, a common practice due to lack of storage space. Errors in ARV data often resulted from haphazard data abstraction from paper records used for clinical care. Sites that used rotating personnel for data abstraction, such as care providers, medical students, residents, and other trainees, appeared to have higher rates of ARV data errors compared with those that assembled focused and well-trained teams. Error rates did not appear to be associated with the level of experience of the local data management team or with the presence of a data center. We currently are performing additional studies to better understand reasons for data inconsistencies.

The audit functioned as a useful data quality control for both the data coordinating center and the participating sites. It allowed the DCC to identify and resolve weaknesses in submitted data before erroneous data could affect study results and provided sites with a baseline estimate of their data quality. As a result, all published CCASAnet studies use revised site data.

Recommendations

The findings prompted us to recommend many of the same quality improvement interventions for each site:

- *Standardize data abstraction, database entry procedures, and personnel training to reduce variability in data quality.* The audit team observed avoidable errors like improperly selected laboratory dates (laboratory results should be paired with the date the sample was drawn rather than the date it was processed or the date of the finalized

laboratory report) and rounded weight values (i.e. 56.7 kg rounded to 56 kg) that led to overall information loss. Such systematic errors could be prevented by educating data abstractors to follow consistent rules during data collection and entry.

- *Develop structured patient visit forms, either paper or electronic, to encourage consistent provider documentation and reduce the amount of missing information.* Our audits determined that the majority of weight-related errors from three sites were due to missing weight measurements. By replacing the blank sheets of paper these sites used for clinician notes with a printed form that prompted providers to record patient weight in a field, some of these missing values could be prevented.
- *Retain laboratory reports whenever possible to reduce the number of sourceless, unverifiable laboratory values and dates.* Sites D, E, and G did not routinely keep copies of the original laboratory reports in the patient medical record, resulting in code 5 errors for associated values (sourceless data). The audit team compared the study data to handwritten laboratory flowsheets when possible, but these were not true source documents.
- *Revise the procedure for storing dates in the database so that data abstractors can accurately record dates that are only known to month or year precision.* We noted that ambiguous dates in the patient record resulted in frequent date-related errors in ARV information.

Each site positively accepted the feedback and submitted a data quality improvement plan to the DCC. We have not yet formally studied the impact of these quality improvement interventions, but early results suggest that the audit process has led to improved site procedures. Our first follow-up audit found a 50% decrease in the overall error rate, with most of the remaining errors resulting from missing

data (values that existed in the clinical chart but were not entered in the database) rather than incorrect information.

Limitations

Our study had both strengths and weaknesses. The analysis of audit findings used a straightforward error categorization system that required little subjective interpretation, making these findings easy to replicate. Our multi-lingual auditors were confident reviewing medical records in French, Spanish, and Portuguese, and their training allowed them to identify causes of error related to clinical process and data handling. Potential variation was minimized by using the same core audit team during every visit. However, the audit team could inspect only a fraction of records at each site, so although records were randomly selected for auditing, the true error rates may differ from the estimated rates reported here. Furthermore, the audit process evolved as the team gained experience with each successive audit, so audits performed later may have been more likely to uncover errors.

Conclusion

Routine clinical care records are a valuable source of diverse, plentiful, and relatively inexpensive medical data for HIV/AIDS research. Without quality control, however, these data may not be sufficiently complete or reliable for research. Investigators who reuse clinical care data must be proactive in addressing potential quality concerns. We do not suspect that the error rates observed in our cohort are substantially higher than those that would be seen if source-to-database audits were performed in other multi-center HIV cohorts. Indeed, several of our sites have participated in other multi-site cohorts such as ART-LINC, TCHARI, and CHIAC [11,98,99]. We do not claim that the findings of other multi-center observational HIV cohorts are erroneous, but it is difficult to interpret study validity

without a formal assessment of data quality. Collaborative research networks – especially those in international settings – should strongly consider implementing formal audit programs to evaluate the reliability of their data sources, to correct discrepancies in data that have already been collected, and to prevent errors in prospective data collection.

On-location audits require often-scarce resources. In order to minimize costs, many of our audits were performed during visits with other scientific objectives. We are currently developing an electronic audit support tool to help standardize and simplify the audit process, allowing auditors to import an electronic dataset, select a set of records to audit, document their findings in real-time using a pre-defined error taxonomy, and quickly generate summaries of audit results. We also are exploring the possibility of incorporating data quality self-assessments as a formal component of our data quality control procedures. With such an approach, sites can perform self-assessments of data quality, which may permit the DCC to reduce the frequency of external audits. Under certain conditions, audit results can be used to statistically adjust estimates based on the original, error-containing data [100]. A data audit should be viewed as an important tool for improving the quality of data, the validity of associated study results, and the reliability of future data collection procedures.

CHAPTER IV

INVESTIGATING PERCEIVED REASONS FOR DATA QUALITY VARIATIONS

Introduction

Meaningful clinical research results derive from complete, accurate, and reproducible study data. In highly regulated biomedical research like interventional clinical trials, the quality of such datasets is assessed routinely through source document verification [52,90]. These audits verify a study's compliance with Good Clinical Practice (GCP) standards and produce quantifiable measures of the accuracy and completeness of research data [1]. Unfortunately audits conducted after data collection is complete often focus on numerical estimates of the error rates in finalized datasets, which rarely leads to a deep exploration of the factors contributing to error [90].

Source verification data audits play an important role in CCASAnet. Between April 2007 and March 2008, members of the network's data coordination center (DCC) visited CCASAnet sites to conduct source verification audits of data submitted for two studies of outcomes in HIV-positive, antiretroviral treatment-naïve persons. The DCC adapted audit techniques and error coding systems used to ensure GCP compliance in controlled clinical trials conducted by the European Organization for Research and Treatment of (EORTC)[101].

During the audit, auditors from the DCC recorded errors of three types: (1) measurements in the submitted research data that did not exist in the clinical record, (2) research data that conflicted with clinical notes, and (3) significant content in the clinical record that was not included in the research dataset. After the audit, team members reviewed the audit findings, classified the errors, omissions, and inconsistencies, and calculated error rates by data category. Auditors calculated an overall error rate per

site, an error rate for each variable type such as patient weight or $CD4^+$ lymphocyte count (CD4), and percentages of incorrect, missing, and unverifiable data.

Such audits provided detailed numerical estimates of the data quality at each site, as described in Chapter III, but were not designed to investigate the causes of observed errors, particularly in variables that proved to have high error rates, such as laboratory data, weight measurements, and antiretroviral data. We decided to revisit the patterns of error detected during on-site audits by surveying site personnel about the perceived reasons for errors in the research data. This study investigated the causes of data quality differences across the sites through an analysis of responses to both general and site-specific survey questions.

Methods

Subjects and Settings

The study population included all employees and volunteers who were responsible for developing, creating, and maintaining sources of data at CCASAnet member sites. Potential participants were identified by local project coordinators and included physicians, nurses, pharmacists, research directors, data managers, and data entry personnel. Our target sample included at least three participants from each site and 21 or more people in total, though we encouraged local site coordinators to involve as many personnel as had worked on the CCASAnet datasets. The study investigators were blinded to the identities of the survey respondents since the site coordinators, not the study investigators, identified potential participants. The Vanderbilt Institutional Review Board approved the study.

Survey Design

We developed a multi-part survey that included general questions about the participating site and its data collection practices, and site-specific questions based on quantitative results from the on-site audits. All multiple choice questions were in English, but study participants were encouraged to complete the free text portion of the survey in the language in which they felt most comfortable. The survey questions did not solicit person-identifying information. Participants were asked to identify the country their site was located in, but not their workplace or role in the organization.

The site-specific questions focused on three classes of data: weights, laboratory results including HIV viral load, CD4, and hemoglobin (where available), and antiretroviral drug regimens. Sites with error rates greater than 10% in any of these variables, according to the preliminary results of the cross-site audit comparison, were asked to describe potential causes of error. The survey included additional, targeted questions if the majority of errors were due to incorrect information, missing source documents, or data that existed in the clinical record but were not included in the study dataset. Sites with error rates less than 5% in any data category were asked to describe what data collection processes they employed to produce better quality results. Table 6 shows which specialized questions were posed to which sites. Sites are labeled A-G to preserve anonymity.

Each question set was implemented in REDCap Survey, a secure, web-based application designed to support research data capture [102]. A link to the custom survey was e-mailed to coordinators at each CCASAnet site, who encouraged individuals involved with the data collection and preparation to complete the survey. We invited study subjects who completed the survey for a free raffle to win one of two iPod Shuffles. Participation in the iPod raffle was optional and entries were not associated with survey responses. We sent two follow-up, reminder emails about the quality survey to coordinators at each site. The survey was distributed in March 2009 and closed in June 2009.

Table 6: Questions included in data quality surveys administered to sites A-G

	Sites						
Topics of survey questions	A	B	C	D	E	F	G
Why are weight values…							
correct						x	x
incorrect	x						
missing		x		x	x		
Why are laboratory measurements…							
correct		x	x		x	x	
incorrect	x						
missing		x	x				
without source documents				x			x
Why are antiretroviral drugs/dates…							
correct			x	x		x	
incorrect	x	x			x		x
missing		x			x		x

Data Analysis

We exported the survey results from REDCap in comma-separated-values format and tabulated multiple choice answers in Excel. Two researchers separately reviewed the free text responses for potential themes. We selected key recurrent themes through iterative analysis and group discussion of our findings. Responses were analyzed in the language in which they were written. For this narrative, the researchers produced translations that were reviewed by fluent speakers. Responses in English by non-native speakers were edited for correct spelling and basic grammar where necessary.

Results

The DCC received eighteen completed surveys, representing six of the seven HIV clinics participating in CCASAnet. Site F never responded to our requests to participate. During the survey period, Site D was employing only two individuals who had been involved with data collection and processing for CCASAnet and therefore could not meet the target of three participants per site. Site A

also submitted only two surveys. We received three to four completed surveys from the remaining sites. The survey response rate, given a target of three responses per site, was 16/21 surveys (76%).

General Questions

Survey participants at two-thirds of the sites reported three to five clinical and data management personnel collected the research data that were submitted to CCASAnet. Data gathering teams were larger at Site B (6-8 people) and Site G (8-10 people.) At all sites except Site E, the majority of on-site data abstraction personnel were clinicians, particularly doctors and nurses, rather than data entry personnel. All sites had data managers or data entry personnel who entered data into the database. The reported characteristics of each site's data abstraction and data entry teams are listed in Table 7 and Table 8.

Table 7: Groups involved in data abstraction, as reported by sites A-E, G
This table displays responses to the survey question, "at your site, what groups collect research information from clinical records?" The number of responses in each category from sites A-E, G, is indicated in the table cell.

Personnel involved in abstracting data from medical records	**# responses at Site**					
	A	**B**	**C**	**D**	**E**	**G**
Doctors	2	4	2		2	3
Nurses		4	2	2		2
Personnel hired for data entry			1		4	
Data managers	1		2			1
Clinic administrators						
Medical students			1			
Pharmacists			1			
Other						

Table 8: Groups involved in research data entry, as reported by sites A-E, G
This table displays responses to the survey question, "at your site, what groups enter information into the database?" The number of responses in each category from sites A-E, G, is indicated in the table cell.

Personnel involved in entering data into the research database	# responses at Site					
	A	B	C	D	E	G
Doctors			3		1	
Nurses			3			2
Personnel hired for data entry	2		1		4	3
Data managers	1	3	3	2		1
Clinic administrators						
Medical students			1			
Pharmacists	1		3			
Other		1	2			

Weight Measurements and Dates

We surveyed four of six sites about causes for incorrect or missing patient weight data. All survey participants at three sites indicated their clinics did not measure weight during every patient visit, particularly for "patients who attend regularly and don't have clinical changes." However, the reasons why clinicians did not assess weight varied by site. Respondents from Site E noted that clinicians "do not judge [weight to be] an important measurement." Furthermore, at this site, weight was a culturally sensitive topic and "was not considered a key variable for… studies." The structure of the site's research database reflected the perceived unimportance of this measurement, as there was only "a place to enter baseline weight but not other weights."

In contrast, patient weight at Site B was not assessed routinely because the clinic had no reliable scales. There was "a lack of material resources -- many broken scales, all in disrepair, and only a few of them (one scale for more than 20 outpatient departments.)" The participant noted that "this makes the practice of measurement infrequent or even discredited, as it is not possible to rely on the anthropometric instrument in question." Respondents from Site A speculated that weights were

measured but never recorded in the clinical record, while participants from Site D were unsure why the data were missing.

The remaining two sites (C, G) captured accurate weight data for ≥95% of cases. One participant described measuring and recording weight as an essential part of the patient visit: “In each medical encounter during the admission of the patient, a paramedic is charged with measuring the current weight, blood pressure, and pulse pressure, and recording those in the clinic file.” They also described having multiple opportunities to capture the same data. If the nurse failed to measure weight during intake, “it is recorded at the same time as the physical examination, by the doctor.” This form of redundant data capture was also effective at reducing the frequency of errors and missing data at site C, where weight measurements were recorded in the chart and also entered in “real time [in]to the electronic medical record (EMR) at each clinic visit.”

Laboratory Results and Dates

We queried five sites about problems with laboratory data quality, primarily missing source documents and missing CD4, viral load, and hemoglobin laboratory measurements and dates. When laboratory values such as baseline viral loads were missing, it was often because the site lacked the resources to request the test (Site C) or clinicians displayed a preference for clinical diagnosis of “patients presenting for care at advanced stages of disease.” or “laboratory problems that severely [delayed] the result.” (Site B)

One of the greatest concerns during the audit was that many sites submitted extensive lab test results for patients in their research dataset, but the audit team could not find corresponding information in the clinical records. At two sites, D and G, over 10% of laboratory values had no source document. Respondents noted that “paper laboratory slips are easy to lose,” but described different

factors that contributed to missing source documents at each site. At Site D, laboratory results were "most often provided in the form of lists," which could be large paper packets that were difficult to attach to the medical record. As a result of the printed format, laboratory documents were "not routinely included in the file," leaving clinicians responsible for transcribing information "from the lab slip into the medical record." At Site G, on the other hand, laboratory results were recorded on small paper slips and the primary cause of missing source documents was a lack of shelf space in the record room and "lack of space in the file." In response to space concerns at site G, "the test results are transcribed into the clinical file (onto the summary sheet for tests) and the laboratory reports are eliminated." Failure to transcribe the data before disposing of the source document resulted in additional missing information. "We don't have space to keep all laboratory slips," wrote one participant, and "if the patient wants the lab printout we give it to them."

However, laboratory data had the lowest overall error rates of any data category on the survey. Such data were rarely incorrect except at Site A, where a typographical error resulted in several database records being paired with incorrect lab values during data export. The attitude of many sites towards CD4 and viral load (VL) results was very different than weights. "Doctors considered these data… important for patient care." Follow-up visit forms often had a "special place to enter CD4 and VL" to promote structured data capture. The research database maintained by Site B received direct data exports from the laboratory system.

Antiretroviral Regimens and Dates

The GCP audits revealed that most sites faced difficulties collecting complete and accurate pharmacy data. The survey queried four sites with error rates >10% about reasons for quality concerns with antiretroviral regimens and dates. The three sites with missing medication data agreed that "data

abstractors sometimes miss old or short ARV regimens" and old treatments at other clinics or "with one or two drugs instead of full HAART" were easily overlooked during data abstraction (Sites B, E, G). Site G also noted that because of insufficient data personnel, "updating the database is not a fast enough process to keep the data up to date."

Respondents from all four sites explained that regular delays between the dates when a physician prescribes treatment, the government approves the treatment, and the pharmacy dispenses the drugs result in multiple, conflicting dates that are recorded in the patient chart, and data abstractors are often unaware which one to use. At Site E, there is "lag time between prescription date (as written in the clinical record) and the actual provision [of medications] by the federal government" and at Site G, there is a "gap between the prescription of the therapy, the petitioning of the ministry of health for the change, and finally the authorization of the change."

The two participating sites with <5% error in ARV data pointed to collaborative work with their local clinic pharmacies as a reason for higher quality. Both sites benefitted from having access to pharmacy records as a secondary source of information. At Site D, "this information is duplicated, is in the files and the database of the pharmacy." Similarly, the "history of change in regimen [at Site C] is kept at three places: the physician EMR [electronic medical record], the pharmacy EMR, the paper chart," providing additional sources of information that helped reduce the number of missing and incorrect values. Unfortunately, Site B's ARV data quality was tainted by pharmacy data that stored incomplete histories of patient medications: there was another source of ARV data (pharmacy), which did not have a log of regimen changes." Sites whose ARV data was imported from a reliable pharmacy system had fewer medication errors. At Site C, "regimen records are entered directly into our database by Pharmacy personnel."

Discussion

According to the audit findings, every participating site encountered difficulties in collecting complete and accurate research data in one or more of the three categories: weight, laboratory, and pharmacy data. The responses of survey participants revealed that different categories of data have errors for different reasons. Weight data were often missing because clinicians assessed patient weight inconsistently, but the causes for infrequent measurements ranged from provider perception of the importance of weight measurements to broken scales. Auditors flagged many laboratory values because of absent source documentation, but responses revealed different reasons at affected sites, including oversized laboratory printouts, a lack of shelf space in the record room, and the practice of giving lab reports to patients.

This survey-based approach generated diverse responses from site personnel; however the written format may not have elicited the greatest detail and variety in responses. Indeed, the audit team reported hearing of other reasons for error that were not mentioned by survey respondents, such as hospital records archives that were reluctant to release records for research, or hired medical students who produced patchwork data abstraction of multi-volume records. Clinicians and data management staff might have been more likely to divulge such information during spoken, face-to-face interviews.

The survey provided insights into approaches to potentially reduce errors in research data abstracted from patient care records. It may be possible to avoid errors in weight data if there are multiple opportunities to measure weight during the patient visit and this measurement is made a clinic routine; in laboratory data, if lab slips are kept with the patient record, not discarded or data are imported directly from the laboratory system; and in ARV data, if data are compared to pharmacy records during data abstraction. However, there is no single data collection method or form capable of resolving all inconsistencies in data quality: the causes of erroneous data are related to both the nature

of the data and the culture of the workplace. By understanding the specific barriers to complete and accurate data collection in different settings, we may be able to effect more rapid, customized, and culturally appropriate changes that improve the reuse of routine patient care data for research.

CHAPTER V

DEVELOPING A COMPUTER-ASSISTED TOOL FOR SOURCE VERIFICATION DATA AUDITS

Introduction

Source verification audits are the gold standard for assessing data quality in clinical research [103]. During such audits, external auditors compare research data to the original documentation of patient care, which may include paper clinical charts, laboratory reports, or the contents of electronic medical record and laboratory systems at the study sites. Protocol-driven studies such as clinical trials often engage teams of clinicians and data managers to perform such audits, to ensure the study generates accurate results and follows best practice guidelines required by the U.S. Food and Drug Administration.

Unfortunately, most verification audits of clinical data use paper forms, which have been shown in general to be less effective and efficient than electronic tools [104,105]. Examples of such forms are available online and in audit review books; many are simple grids, with no pre-printed elements from the database and only one column per record [67,106,107]. In such forms, auditors have no space to take notes or record corrected values; they only mark each data element as “satisfactory” or “unsatisfactory”.

Although such paper-based audits are still common in medicine, computer-assisted audit tools (CAATs) have improved the quality of audits in finance, manufacturing, and network security by facilitating more thorough audits, generating more consistent documentation, and saving both time and money for auditors and auditees [108-110]. Specialized CAAT software can aid auditors during many

stages of the audit process, including selecting an audit sample size and methodology, merging and analyzing data, and generating audit reports.

Financial audit software dominates the CAAT market. By the end of the 20th century, virtually every accounting firm conducted audits using commercial analysis suites (e.g., Audit Command Language (ACL) and Interactive Data Extraction and Analysis (IDEA)) or similar proprietary software [111]. Such CAATs assist auditors in downloading data from clients' accounting databases, selecting the audit sample size and audit methodology, conducting cross-database comparisons, and detecting outlying, duplicate or potentially fraudulent transactions [112,113]. The software also provides user support based on guidelines and standards published by professional auditing societies, including the Financial Accounting Standards Board and the American Institute of Certified Public Accountants [108,114,115]. Each single-user ACL or IDEA license provides access to powerful data analysis tools, but also costs thousands of dollars. Less expensive audit-specific software includes TopCAATs, a Microsoft Excel audit plug-in, and Picalo, a Python-based, open-source data analysis and fraud detection toolkit [116,117]. Auditors also use statistical or data extraction software as a CAAT in order to detect anomalous patterns in large datasets. Most multi-purpose CAATs require computer programming skills.

Some CAATs have been developed to assess quality in healthcare. UMTAudit, a large commercial audit suite, handles customized inspections in any industry but can be configured for medical compliance reviews, hospital accreditation audits, and adverse drug event investigations. The application allows auditors to assemble lists of questions to ask site personnel and supports the use of handheld devices for recording information and taking photos during walk-through inspections [118]. The Global Fund to Fight AIDS, TB, and Malaria distributes a different type of CAAT for measuring the quality of aggregate data reported to national health programs and donor-sponsored projects [119,120]. These form sets are available as Excel worksheets, named the Data Quality Assessment (DQA) for

external site inspections and the Routine Data Quality Assessment (RDQA) for internal assessments of program-level quality indicators. Both worksheets require users to input numeric results of data quality audits, but do not provide tools to conduct actual data validation against source documentation. A third type of healthcare-oriented CAAT -- clinical audit software like Auditmaker or PCS Clinical Audit Tool -- uses existing clinical care databases to generate aggregate data about the quality of physician practice, to improve patient outcomes and compliance with standards of care [121-123].

All these audit software packages focus on analyzing an existing, electronic dataset for errors and unusual patterns, rather than facilitating the comparison between the dataset and a physical source document. Indeed, in many accounting and security audits, the electronic database *is* the source document and no other records exist. As a result, these advanced software packages are not helpful for auditing paper source documents. Furthermore, the high cost of tools such as ACL, IDEA, and UMTAudit makes purchasing them unfeasible in resource-limited settings [124].

We believe a CAAT specifically designed for research data auditing can improve the process of source document verification. This project aims to (1) identify the core weaknesses of paper forms when used for clinical data auditing, (2) develop a set of functional requirements for a computerized audit tool, (3) implement the necessary requirements in a prototype audit system, and (4) test the resulting application during six data audit visits at participating HIV clinics. The CAAT desiderata and implementation are described in subsequent sections.

Motivation

Study Setting

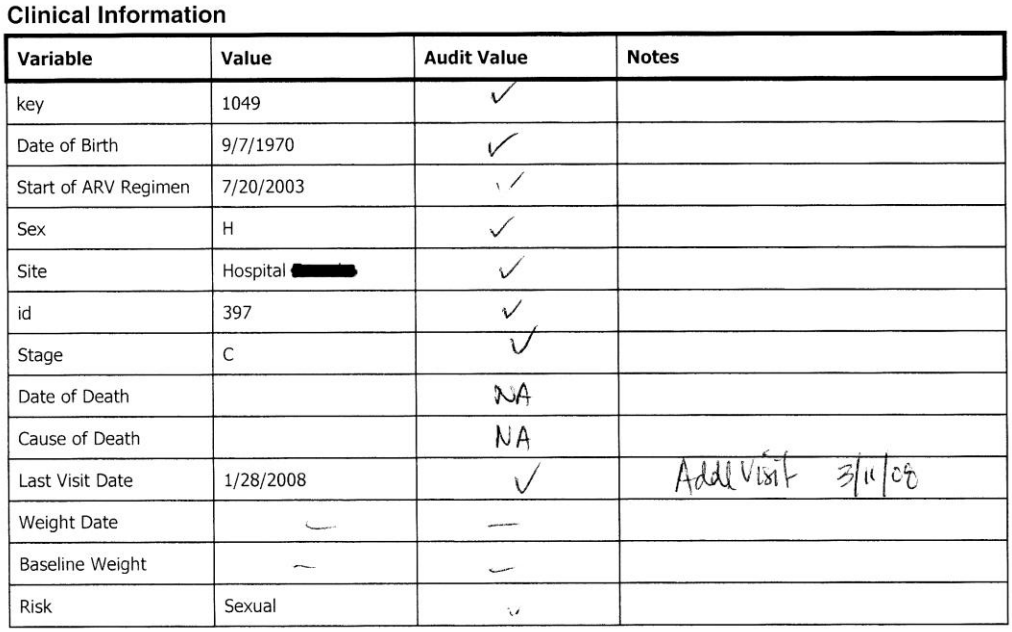
The Caribbean, Central and South America Network for HIV epidemiology (CCASAnet) brings together researchers from Argentina, Brazil, Chile, Haiti, Honduras, Mexico, and Peru to investigate region-wide trends in the HIV epidemic [82]. Participating HIV clinics contribute de-identified, routine patient care data for proposed studies and the separate datasets are merged by the network's data coordinating center (DCC), housed at Vanderbilt University.

To ensure the reliability and completeness of study data, the DCC conducts routine Good Clinical Practice-based audits of the datasets submitted by CCASAnet member sites. During such audits, a small team of clinicians and informaticians visits each member clinic to compare the on-site medical documentation to the contents of the electronic dataset the site previously submitted for analysis.

Paper-based Audit Process

The first CCASAnet audit circuit was completed in April 2008 and included seven baseline site audits and one re-audit. During all eight visits, the audit team used a multi-page paper audit form to record the results of the database-to-clinical record comparison. Data on the form were divided into categories including demographics, clinical visit data, antiretroviral regimens, and laboratory results. The form contained two preprinted items for each data element: the name of the variable (e.g., birth date, date of death, viral load result), and the corresponding value from the site's submitted database. The team used the blank "audit value" field to record whether a data element was present in the source documents and whether the source value was correctly represented in the database. A small notes field

– as well as the margin of the paper – was used to record additional information or possible causes of the error. Figure 7 shows a sample page of a completed audit form.



Clinical Information

Variable	Value	Audit Value	Notes
key	1049	✓	
Date of Birth	9/7/1970	✓	
Start of ARV Regimen	7/20/2003	✓	
Sex	H	✓	
Site	Hospital [redacted]	✓	
id	397	✓	
Stage	C	✓	
Date of Death		NA	
Cause of Death		NA	
Last Visit Date	1/28/2008	✓	Add Visit 3/11/08
Weight Date	—	—	
Baseline Weight	—	—	
Risk	Sexual	✓	

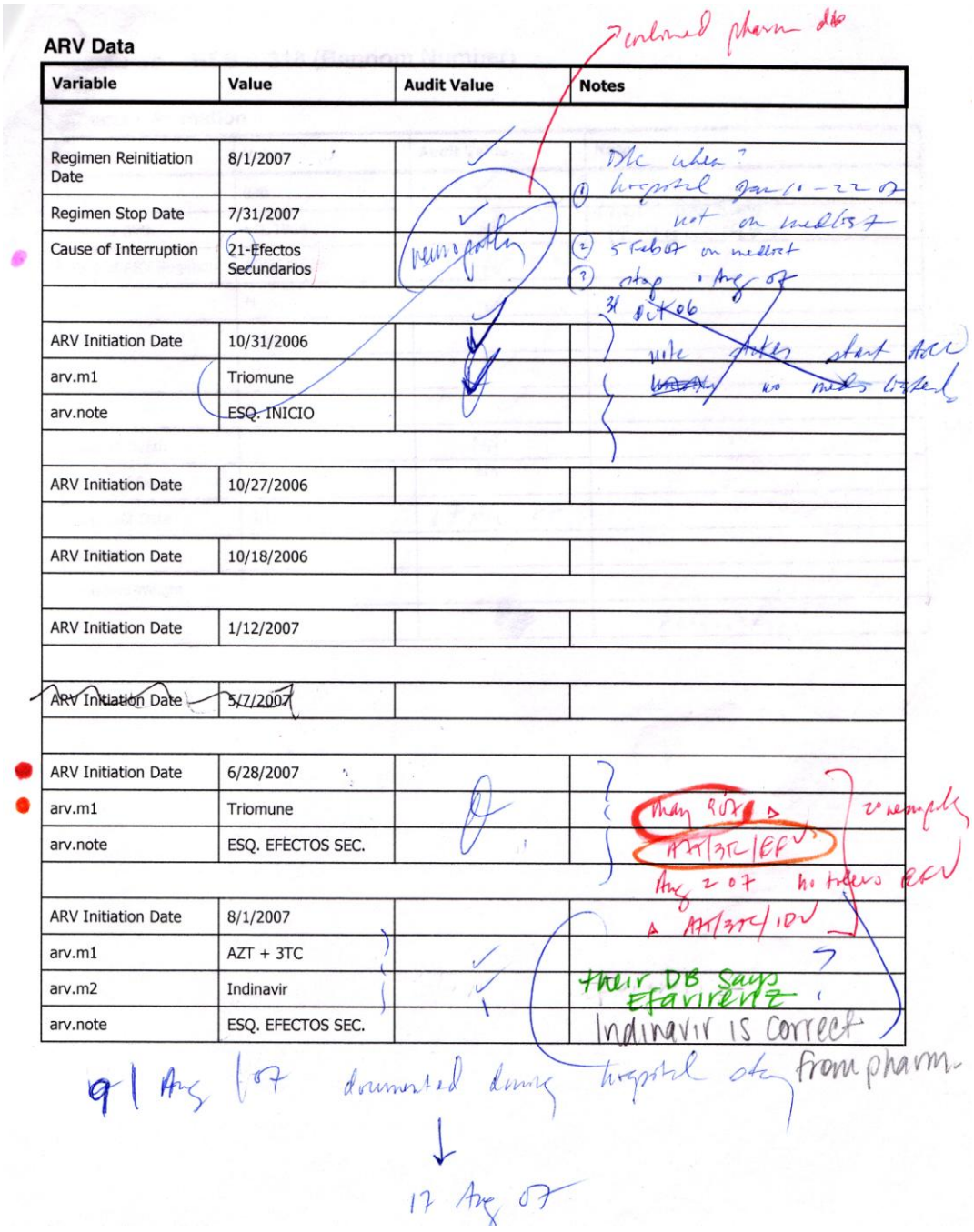
Figure 7: A neatly completed paper form from a CCASAnet site audit
This scanned image shows the first page of a multi-page paper audit form used for on-site CCASAnet audits.

At the end of an audit visit, the auditors presented their preliminary findings during an exit interview with the site investigator and staff. After returning to the DCC, the audit team reviewed the audit forms, categorized and tabulated the data errors, and produced a report describing its findings and recommendations, which was sent to the site for review and comment. The audit team inspected 210 randomly selected records and 4,686 unique data elements during eight audits.

Limitations of Paper-based Audit

CCASAnet's paper-based audit process relied heavily on memory, interpretation, and opinion, and was difficult to replicate and standardize across sites. Although the auditors were not noticeably

frustrated with the paper form during the audit, the loose structure of the paper audit form allowed ambiguity during data collection. When the DCC undertook a reevaluation of the audit findings in mid-2008, the auditors had a difficult time classifying the data discrepancies according to a formal error taxonomy because the original paper forms had required auditors only to describe and correct errors. Of 210 original audit forms assessed during the reevaluation, 30 (14%) contained auditor notes that were partly illegible and 56 (27%) contained underspecified error descriptions that made error coding challenging. Figure 8 depicts such a record with unclearly documented audit findings. Furthermore, the average time between the end of an audit and the completion of the audit report was 101 days, which meant the site data personnel rarely received immediate, implementable recommendations on how to improve data quality.



ARV Data

Variable	Value	Audit Value	Notes
Regimen Reinitiation Date	8/1/2007		
Regimen Stop Date	7/31/2007		
Cause of Interruption	21-Efectos Secundarios		
ARV Initiation Date	10/31/2006		
arv.m1	Triomune		
arv.note	ESQ. INICIO		
ARV Initiation Date	10/27/2006		
ARV Initiation Date	10/18/2006		
ARV Initiation Date	1/12/2007		
ARV Initiation Date	5/7/2007		
ARV Initiation Date	6/28/2007		
arv.m1	Triomune		
arv.note	ESQ. EFECTOS SEC.		
ARV Initiation Date	8/1/2007		
arv.m1	AZT + 3TC		
arv.m2	Indinavir		
arv.note	ESQ. EFECTOS SEC.		

Figure 8: A messy paper form from a CCASAnet site audit
This scanned image shows the final page of a CCASAnet paper audit form. The form contains markings by three different auditors and demonstrates the problems of unclear handwriting, ambiguous documentation of audit conclusions, and the challenge of post-audit error coding.

In post-audit debriefings, the auditors identified several causes for delays in producing the audit report, including difficulties with

- managing multiple audits and reports simultaneously,
- interpreting partially or incorrectly completed audit forms,
- assessing whether observed data discrepancies were clinically meaningful,
- assigning audit codes to different types of errors after the audit,
- consulting with other auditors about error classifications or unclear information,
- reading other auditors' handwriting on completed audit forms,
- interpreting underspecified auditor notes without the presence of the source documents,
- sharing a single set of original, paper audit forms among a team of auditors,
- tabulating errors,
- double-checking other auditors' error tables,
- calculating error rates,
- composing a thorough and detailed audit report, and
- formatting error tables for the final document.

Feedback on data quality is most effective when it is communicated shortly after the audit takes place, but the use of paper audit forms made generating accurate and timely reports a challenge [62]. We hypothesized that a computer-assisted audit tool that replaced the paper forms had the potential to reduce the number of indecipherable or ambiguous audit findings and increase the timeliness and reproducibility of audit results. Our study examined the feasibility of CAAT-based auditing in resource-limited settings by determining the desirable functions of a computerized audit tool, implementing

these features in a CAAT prototype, and testing the resulting application during the second CCASAnet audit cycle.

Key Attributes of an Audit Tool

Through a review of experiences documented during the initial audits, post-audit debriefings, and audit-related discussions with the DCC and CCASAnet sites, we identified five key areas in which a computer-assisted audit tool could improve the audit process: networking, audit data management, standardized error assessments, audit decision support, and results reporting [125]. The audit team's experiences that motivated these requirements are described in the text. The requirements are outlined as desiderata in Table 9.

Networking

Our discussions and review identified an effective CAAT as a networked application that could accommodate multiple, simultaneous users. Paper forms had functioned as an excellent sharing tool during audits. Although each auditor worked independently on a set of records, difficult cases or records with cascading errors often required group review in which the source document and paper audit form were passed around the table. A suitable CAAT needed to facilitate the same real-time communication between multiple auditors. It had to allow collaborative editing of a single copy of the data, as web-hosted applications do, but could not be hosted remotely at the DCC because (1) Internet access was unstable or unavailable at some CCASAnet sites, and (2) long network delays (high latency) between webservers in the U.S. and CCASAnet sites in the Southern Cone would result in a lag-prone user experience. A suitable audit tool needed to take advantage of alternate network structures, such as locally hosted applications and wireless ad hoc networks between auditor laptops.

Audit Data Management

Generating paper audit forms in advance of each audit was a laborious, multi-day task for the audit team and the CCASAnet data manager. If an audit tool could generate web forms dynamically by combining user-designed form templates with pre-formatted datasets, the preparation time would decrease, especially when evaluating the same template at multiple sites. A standard XML data specification would permit auditors to load a copy of the audit data, as provided by the study data manager, before the audit began. A standard for data import/export would allow a copy of the audit results to be delivered to the auditee in an electronic format.

Standardized Assessment of Errors

The paper audit forms prompted auditors only to record and describe errors during the on-site audit; errors were counted and categorized by type during the preparation of the audit report. An effective CAAT would prompt auditors to assess and categorize errors *during the audit visit*, rather than weeks afterward when the source documents were no longer accessible.

Audit Decision Support

Selecting the number and type of records to audit can be a challenge for novice auditors. The CCASAnet audit team consulted a statistician in advance of each audit, but a useful CAAT could provide basic guidance on sample size calculations and selecting records for audit, using selection metrics that have been described in the literature. Contingent on additional studies of error types, it might be possible to incorporate decision support for error classification; a mismatched weight value of 57.5kg in the clinical record vs. 58kg in the database, for example, is likely to be a rounding error of limited clinical significance. This functionality could be useful in labeling complicated errors of drug prescription and

discontinuation. A CAAT with full decision support capabilities could also provide post-audit recommendations based on the numeric audit results.

Results Reporting

The CCASAnet team found that preparing a post-audit report from paper audit sheets required tedious and time-consuming work. Both auditors needed to count the variables in each record, group any recurring errors, double check final numbers, and tabulate the results manually, which delayed preparing the final report. Software support to tally errors and generate tables could allow auditors to present an accurate summary of findings during the exit interview as well as simplify the production of a post-audit report.

All five desiderata are listed in a table on the next page.

System Design

Setting

The network's second audit cycle began in April 2009 with the submission of new datasets for studies of cancer and tuberculosis (TB) in HIV-positive persons. Data abstractors at each CCASAnet site identified cases of TB and cancer in local clinical charts and entered related demographic, diagnosis, and treatment-related data into REDCap, a secure, web-based application for research data entry [102]. REDCap featured separate, multi-page forms for TB and cancer data entry and separate databases for each site. The template for each form could be exported from REDCap as a data dictionary in comma-separated-values (CSV) format, specifying the form fields, their accompanying labels, a coded value set (if applicable), and the appropriate HTML form element (checkbox, text field, and dropdown list). The

de-identified, observational data entered into REDCap could also be exported in CSV format with one line per record.

Table 9: Desiderata for a computer-assisted tool for data audits

Obstacles Encountered during Audits	**Solutions / Desiderata for a Computer-Assisted Audit Tool**
Challenge: Collaboration	**Solution: Networking**
Auditors need to work collaboratively on the same copy of a record.	Real-time Collaboration: networked laptops for auditors, shared databases, web-based systems
Audit sites may have no network infrastructure.	Portable Network Infrastructure: peer-to-peer networking, portable server and router
Challenge: Audit Data	**Solution: Audit Data Management**
Paper audit forms take a long time to prepare and validate.	Import Functionality: one-click import of data and data descriptions (metadata) from research database to CAAT, instant generation of basic electronic audit forms
Copying audit results from paper forms into a spreadsheet for analysis is time-consuming.	Export Functionality: export of audit results into structured data formats (e.g., CSV, XML)
Datasets may contain different medical content (e.g., HIV, Tuberculosis, or cancer data).	Metadata Management: customizable display of data on screen, data dictionaries for special topic areas (HIV, TB, Cancer)
Challenge: Types of Errors	**Solution: Standardized Assessment of Errors**
Errors are not categorized and described clearly on paper forms, making it difficult to analyze and report error types and rates.	Representation of Error Types: categorization of errors, clear operational descriptions of error types, specification of domain of error types (applies to specific variables within the audit record or applies to entire record)
Challenge: Audit Design and Conduct	**Solution: Audit Decision Support**
Auditors are unsure how many records should be audited to produce meaningful results.	Statistical Dashboard: guidance for sample size calculations, identification of grossly problematic records, pre-selection of records via statistical sampling
Auditors may not know which recommendations to offer based on audit findings.	Results-based Recommendations: automated suggestions for quality improvement measures based on classes of errors detected in the data.
Challenge: Analyzing and Presenting Results	**Solution: Results Reporting Tools**
Tallying and tabulating errors by hand is a time-consuming and error-prone task for auditors.	Automatic Error Tabulation: software support for generating tables and graphs
Sites enjoy receiving copies of the individual record audit forms.	Automatic Form Generation: software support for producing printable forms that mirror the format of completed paper audit forms.

The second-round audits were conducted at nine separate locations, including clinic libraries, hospital conference rooms, record review rooms in the hospital archive, and public office spaces. Three locations had reliable, broadband internet access, two had slow and intermittent access, and the remaining four had no internet access.

Implementation

The DCC needed an audit system that could accommodate the cancer and TB datasets submitted by every site, while we endeavored to adhere to the CAAT desiderata we had established. We implemented the *networking* attribute by designing a web-based application that was hosted by an Apache 2.2 webserver running on one auditor's laptop computer [126]. We coded the web application in PHP and implemented dynamic user interface elements using the jQuery JavaScript library [127,128]. The audit data were stored in a parallel MySQL 5 database [129]. Users connected to the webserver using available local area networks at the CCASAnet clinics. When Internet access was unavailable, the audit team established a computer-to-computer network between auditor laptops. After the completion of the audit, we relocated the application files to a secure Vanderbilt webserver, so all auditors could access the electronic audit records while preparing a post-audit report.

The REDCap system simplified our approach to *audit data management* since we received each site's cancer and TB dataset in a standardized, CSV export format. We had previously encoded the structure of our cancer and TB case report forms in order to generate data entry forms in REDCap for interested sites. We developed a metadata import function that reused these field descriptors to construct the layout of our audit forms. Appendix B provides examples of our CAAT data import format and data dictionary, both based on the REDCap data specification.

To maintain comparability with our previous audits, we implemented the European Organization for the Research and Treatment of Cancer (EORTC) *error classification system* described in Chapter III, where code 1 represents correct data and codes 2-5 represent minor errors, major errors, missing data, and sourceless data, respectively. When auditors flagged data discrepancies, we prompted them to code the error type based on a dropdown selection from the EORTC list.

We chose not to implement *audit decision support* for this prototype since we already received our lists of records to be audited from the biostatistician who had designed our previous audits. We acknowledge that automatic tools for selecting audit records would be a valuable CAAT feature, but we would not have had occasion to test it using our established audit methodology. We also lacked the necessary root cause data to build an audit decision support module that could offer quality improvement recommendations based on numeric audit findings.

We automated the post-audit tabulation of errors using a *results reporting* system that calculated the frequency of error types as well as overall error rates. Since our auditees were accustomed to paper audit forms, we also generated a printable document displaying our audit findings in the familiar format.

Application Walkthrough

We present the reader with a descriptive walkthrough of our prototype CAAT to illustrate our implementation choices and the application's functionality. All screenshots presented in Figure 9 though Figure 15 were taken in Mozilla Firefox 3.6.

Our prototype CAAT had five core modules that were linked from the application's homepage (Figure 9): data upload, record selection, record auditing, report generation, and application configuration.

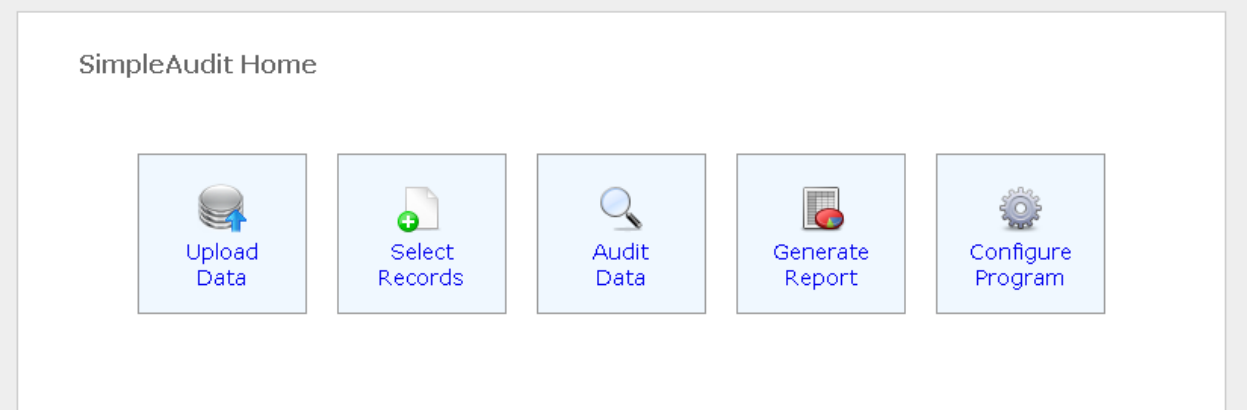


Figure 9: Screenshot of the homepage of SimpleAudit, a prototype computer-assisted data auditing tool
This screenshot shows the homepage of our data auditing application in a web browser. The five core modules are displayed: data import, audit record selection, auditing, report generation, and application configuration.

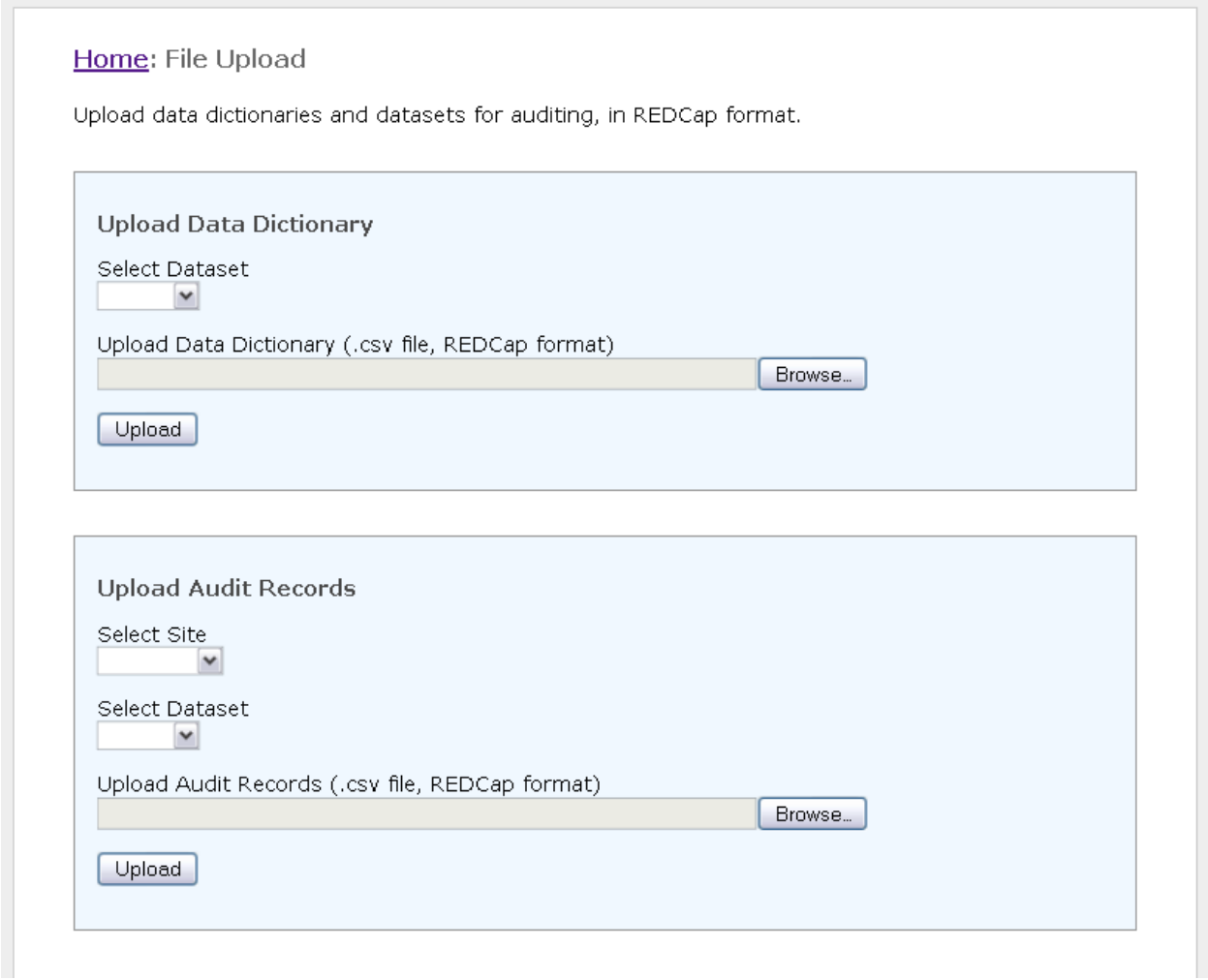
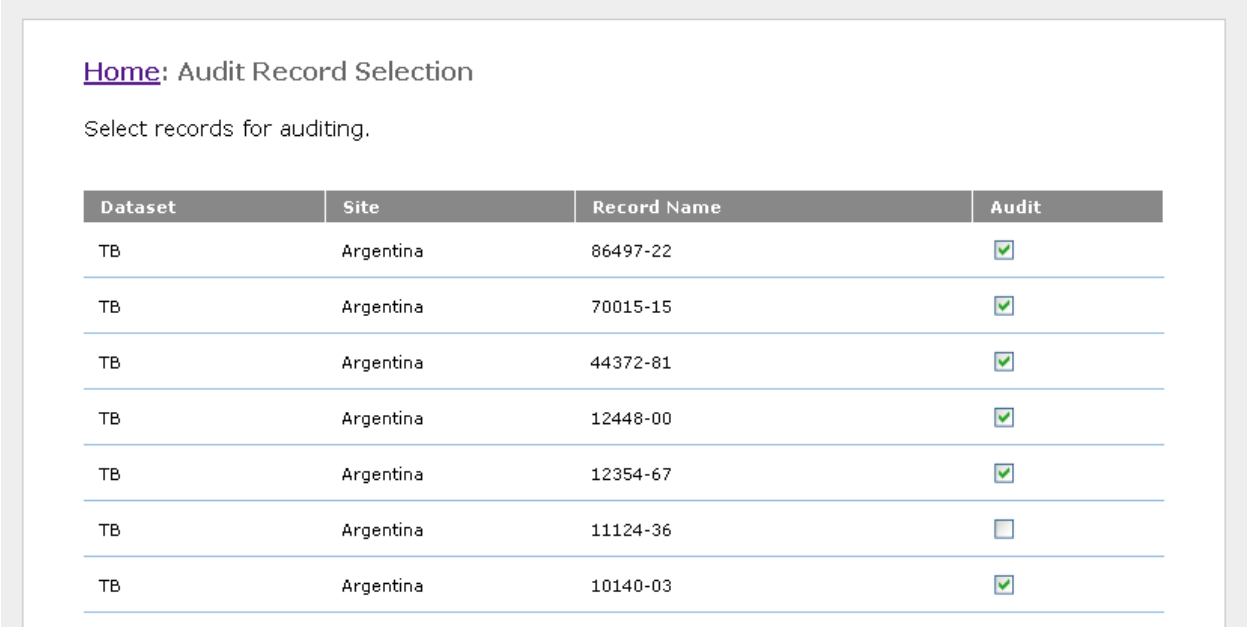


Figure 10: Screenshot of the data import screen
The Data Import screen has forms for uploading two content files: a dataset from which records can be selected for auditing and a data dictionary describing the structure of the record file.

The Configuration module handled user and audit setup, allowing the audit manager to create new audit events, label imported datasets, register new users, assign passwords, and track logins.

The Data Import module transferred research datasets into the audit system. The upload form, shown in Figure 10, prompted the user to upload the data dictionary and dataset as exported from REDCap. The data dictionary determined the labels for the variables and order of the pages of the auditing form. The application then parsed the data and metadata files and stored the contents in several entity-attribute-value database tables, where each form field and data element was assigned a unique ID. Incorrectly formatted research datasets generated alerts that halted the import process and prompted the user to manually repair the data files. In the last step of pre-audit preparations, the auditor designated a subset of records for on-site review. The Record Selection screen (Figure 11) displayed a list of all imported record IDs and allowed the user to choose specific records by ID number. We selected records according to a list provided by our biostatistician, but if we had implemented decision support for record and sample size selection, it would have been incorporated in this module.

Once on-site, auditors tested the audit support capabilities of the CAAT. The home screen of the Auditing module, seen in Figure 12, displayed a list of all the records that had been selected for review, along with the name of the dataset being audited and the method used to select the record. Once auditors signed off on the review of a record, the link to that record was moved from the “Incomplete Records” to the “Completed Records” list.

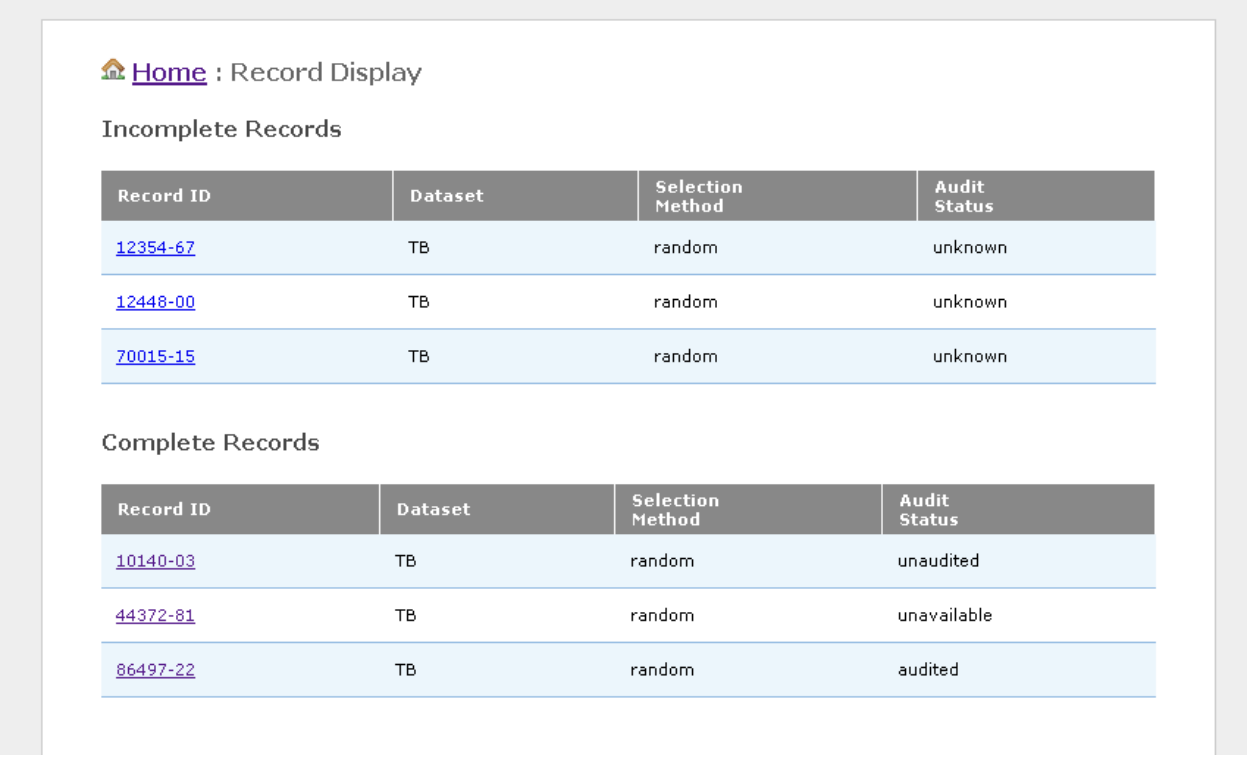


Home: Audit Record Selection

Select records for auditing.

Dataset	Site	Record Name	Audit
TB	Argentina	86497-22	☑
TB	Argentina	70015-15	☑
TB	Argentina	44372-81	☑
TB	Argentina	12448-00	☑
TB	Argentina	12354-67	☑
TB	Argentina	11124-36	☐
TB	Argentina	10140-03	☑

Figure 11: Screenshot of the audit record selection page
Auditors choose records for auditing (from a list of all uploaded records) using the page-right checkboxes.



Home : Record Display

Incomplete Records

Record ID	Dataset	Selection Method	Audit Status
12354-67	TB	random	unknown
12448-00	TB	random	unknown
70015-15	TB	random	unknown

Complete Records

Record ID	Dataset	Selection Method	Audit Status
10140-03	TB	random	unaudited
44372-81	TB	random	unavailable
86497-22	TB	random	audited

Figure 12: Screenshot of the listing of records to be audited
All the records that have been selected for auditing are listed on the Record Display page. Records move to the bottom list once auditors mark that record “complete.”

Auditors interacted primarily with the main audit screen pictured in Figure 13. Each audit record had a customized, dynamically generated audit form. The ID of the audit record was displayed prominently at the top of each page, followed by hyperlinks to the different pages of the audit form.

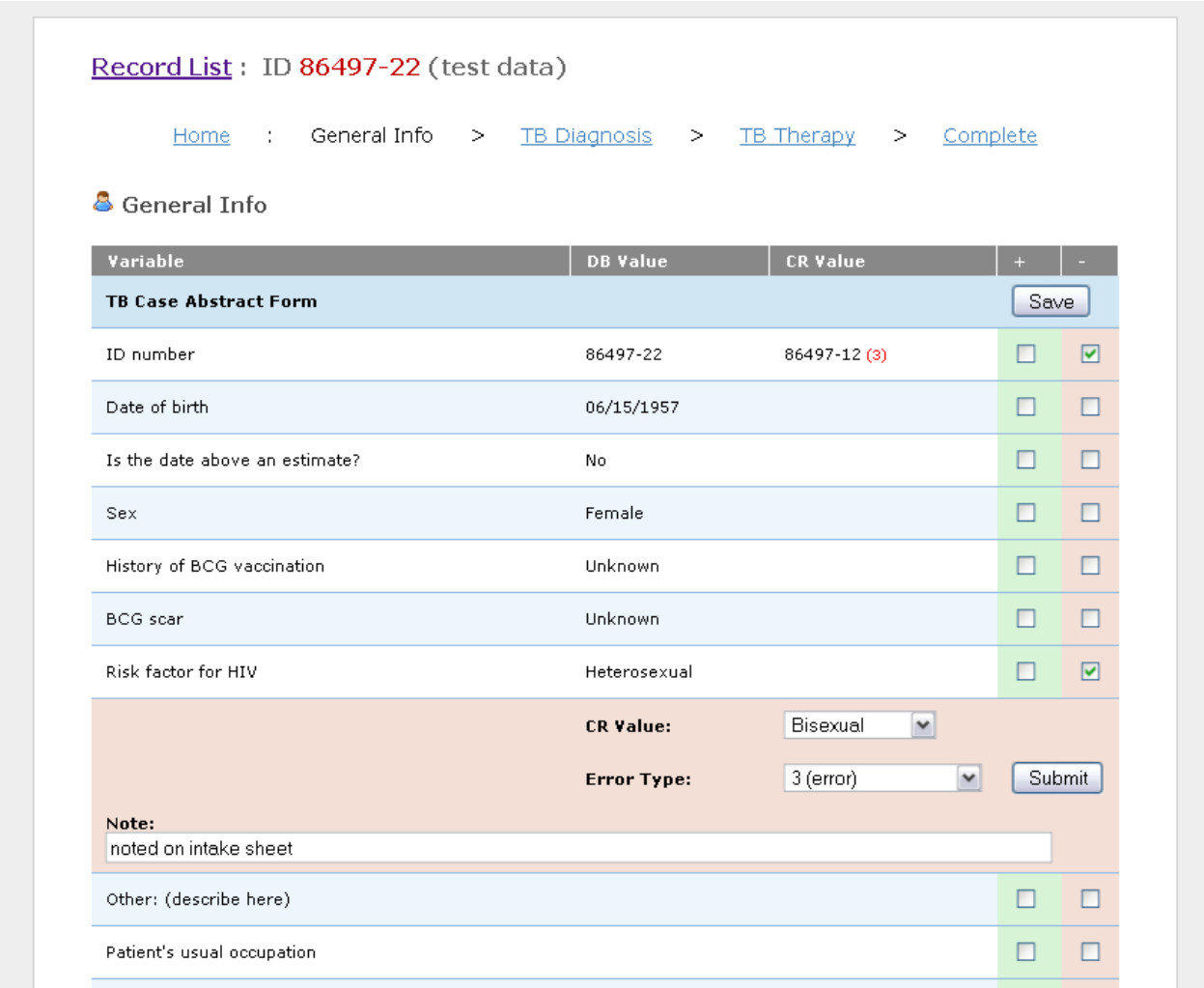


Figure 13: Screenshot of the data audit interface
The computer assisted audit tool provides a custom, dynamically generated web form for each audit record. The image above depicts a partially audited example record, whose ID number, 86497-22, is shown in bold red font at the top of every page of the form. The form above has three pages: General Info, TB Diagnosis, and TB Therapy, followed by the record completion page. In the example above, an auditor has noted an error for the variable "Risk factor for HIV," by checking the box in the red "error" column, indicated by the '-' heading. This action causes an extra row, shaded in red, to expand below the checkbox. The fields in the extra row allow the auditor to record additional information about the error. In the example above, the auditor has also marked "ID number" as incorrect. The corrected value and the error code are shown in the "CR value" column.

The audit interface was a simple five-column form listing the label for the data element (e.g., Date of birth, Sex, Risk factor for HIV), the value the site had submitted to the DCC ("DB value"), a space for the value the auditors identified in the clinical record, if different ("CR value"), and two checkboxes to mark a correct value ("+") or an error ("-"). When an auditor checked the error column, the web form expanded to add a temporary row for entering error details, shown with a red background in Figure 13. This sub-form prompted auditors to enter a corrected value and code the type of error according to the classification developed by the EORTC, described in Chapter III. The error form also contained a text field for auditor notes.

Once an auditor finished verifying all necessary data elements in a given record, he marked the record as "audited" using the record completion form shown in Figure 14.

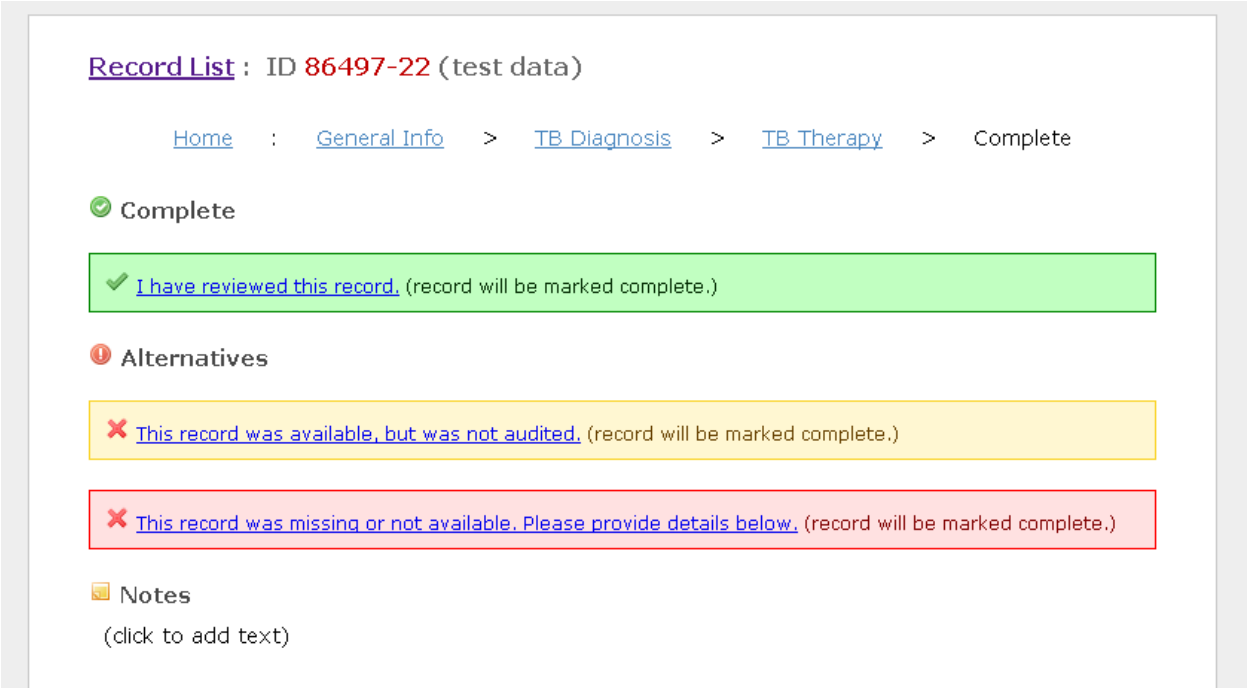


Figure 14: Screenshot of the record completion screen
Once an auditor has completed his review of a record, he marks the record complete by selecting the green "complete" box pictured above. Alternate selections include "unaudited" for records that were available but not audited (yellow box) and "unavailable" (red box) for records that the site could not locate or provide.

Fully reviewed records were marked as “audited.” Alternate completion labels included “unaudited” for records the team received but lacked time to audit and “unavailable” for records the site could not locate or provide during the audit period. Selecting any of these three options moved the record to the “Completed Records” section of the record list. The final Reporting module assisted auditors with summarizing the audit findings and preparing a report for the sites. The CAAT automatically tabulated the number of correct, incorrect, missing, and no source document errors from records marked “audited” and highlighted error rates over 10% in red, as shown in Figure 15. It also generated printable forms, shown in Figure 16, that displayed the audit findings in the format of the original paper audit forms (shown previously in Figure 7 and Figure 8).

Fields on TB Form	Total Variables Reviewed	Correct	% Correct	Field Mismatches	Missing/ Missed Data	Data with no Source Document	Error Rate
TB Case Abstract Form							
ID number	21	21	100%	0	0	0	0
Date of birth	21	21	100%	0	0	0	0
Is the date above an estimate?	21	21	100%	0	0	0	0
Sex	21	21	100%	0	0	0	0
History of BCG vaccination	18	18	100%	0	0	0	0
BCG scar	18	18	100%	0	0	0	0
Risk factor for HIV	21	21	100%	0	0	0	0
Other: (describe here)	8	6	75%	0	2	0	0.25
Patient's usual occupation	16	1	6%	0	15	0	0.94
Patient was household contact of a TB case	18	18	100%	0	0	0	0

Figure 15: Screenshot of the audit summary report
The audit summary table displays a list of the audited data elements and a numeric summary of the audit results. The report lists the number of times each type of variable was audited, the number and percentage of correct data elements, the number of data elements that were incorrect, missing, or lacking source documents, and the variable-specific error rate. Table cells with error rates higher than 10% are highlighted in red in the far right column.

Record ID: Research ID (test 1)

Chart Notes:

In 2004-Nov-18 visit note: mention of carcinoma in situ, planned surgery in December. But this record ([redacted]) is not in Cancer database. We should probably review the record and see if it should be added to the Cancer dataset.

Variable	DB Value	CR Value	Error Type	Note
TB Case Abstract Form				
ID number	[redacted]		1	
Date of birth	5/3/1975		1	
Is the date above an estimate?	No		1	
Sex	Female		1	
History of BCG vaccination	Unknown		1	
BCG scar	Unknown		1	
Risk factor for HIV	Heterosexual		1	
Other: (describe here)				
Patient's usual occupation		cadera	4	cadera, maybe cadena?
Patient was household contact of a TB case	Unknown		1	
PPD+ (>= 5mm)	Unknown		1	
Date of PPD result				
Is the date above an estimate?	No		1	
INH prior to TB	Yes	No	3	Patient had prior exposure to INH during previous TB tx (in 2000), but this question is about INH treatment (alone) before TB

Figure 16: Automatically generated audit forms
The CAAT automatically generated these printable forms based on the imported data, imported form template, and auditor findings.

The final system contained five core modules that supported users during the preparation for, conduct of, and reporting of a clinical research data audit. We implemented four of our five proposed CAAT desiderata in this prototype.

User Experience

The application configuration presented no problems; we hosted the webserver on an auditor laptop running Windows XP. All sites except Site C had submitted data through REDCap and these data and metadata files presented no import problems. Site C's dataset generated minor formatting alerts that prompted us to correct the data file contents. Our estimates suggest this application decreased the average audit preparation time to less than one hour (from greater than six hours).

The auditors were pleased with the physical convenience of using a CAAT as it freed us from transporting both a laptop and up to three pounds of paper audit forms during travel and site visits. Desk space on-site was also limited, and auditors favored placing a laptop on the table rather than paper forms that occasionally vanished under piles of clinic charts. Electricity for charging laptop batteries was available at every audit site. (On the other hand, the auditors did find the CAAT was not as useful as the paper forms for swatting mosquitoes.)

The web application had a low response time when handling auditor requests over most clinics' local area networks, but difficulties arose in locations with unreliable internet access. The computer-to-computer network was unstable, sometimes requiring several computer reboots before our laptops would connect. We discovered the ad-hoc network configuration was not robust enough to handle more than two auditor laptops; with three simultaneous users the poor response times from the webserver made the application unusable.

Typed text eliminated illegible auditor notes and the CAAT promoted immediate rather than post-audit coding of all data errors. Auditors used the results reporting module to present numeric audit results during the exit interview, generate tables for the audit reports, and generate printable results forms. These forms improved the transparency of CCASAnet audits by allowing site personnel to see our conclusions for each record and variable.

Our informal evaluation indicated that this prototype application decreased the average audit preparation time to less than one hour (from more than six hours), increased auditor compliance with documentation and coding, eliminated the transcription of audit results, and provided sites with more detailed audit results.

Discussion

Summary of Findings

When an audit process lacks standardization, as our original paper-based audits did, different auditors may produce different audit reports given the same source record. Without standardized quality measures, the changes in an organization's data quality cannot be compared from year to year, nor can audit results be compared from site to site. We described five desirable attributes for a flexible audit support tool designed to simplify and standardize auditors' work, including multi-user support through networking, file formats for importing and exporting audit data, standard error assessment protocols, audit decision support, and tools for calculating and displaying audit results.

Our prototype system demonstrated the significance of these key attributes as well as the feasibility of implementing them in a simple web-based application. We tested our CAAT during audits of twelve cancer and tuberculosis datasets in nine locations and found the system drastically improved the audit process by decreasing preparation time from over six hours, on average, to less than one hour; enforcing consistent, on-site error coding for all records; and eliminating difficulties reading other auditors' completed forms, as well as manual error tallying for the post-audit report.

Limitations

The CAAT recommendations we describe stem from audit experiences at HIV clinics in Latin America and the Caribbean. The perspective of a single international research network, however, may limit the diversity of experiences that informed the five suggested requirements. Indeed, for single-user audits, a desktop application or other non-networked solution might present less complexity. Considering multidisciplinary audit teams are more common and effective [130].we believe these recommendations represent a necessary though perhaps not sufficient set of features needed for a successful paperless audit tool.

Additionally, our CAAT was limited in both its implementation of our proposed key attributes and its general applicability. We elected not to implement the fourth attribute, Audit Decision Support, as it was redundant given our audit circumstances and we were unable, therefore, to evaluate its usefulness in the field. We also did not write installation instructions or support documentation as the system was a prototype. A full-fledged application should include such support tools and documentation.

Our approach to importing audit data and form designs as REDCap CSV files saved several hours of preparation time as we had already designed the electronic form templates to create REDCap data entry interfaces for CCASAnet sites. Other users wishing to audit datasets for which no REDCap form templates exist would need to create them and would therefore not benefit from these time savings. Nevertheless, the REDCap data and metadata files are built using a straightforward design and coding process (as shown in Table 15 and Table 16 of Appendix B), and require no expertise beyond typing a list of variable names, labels, and form element codes into an MS Excel spreadsheet. We believe designing one such template for several audits of identically formatted data is much faster than creating new paper audit forms for each site.

We also recognize the effectiveness of the REDCap-motivated audit interface is dependent on the structure of the audit data. As we determined during the initial paper-based audits, correcting errors in antiretroviral regimen data often requires auditors to rearrange the order of regimens and insert new fields to accommodate additional data points. The TB and Cancer datasets we reviewed using the CAAT contained no variables that were as order-dependent as ARV regimens. Such data would require extended user interface option, possibly with Javascript widgets that allowed drag-and-drop duplication and reordering of medication regimens.

We were unable to conduct a quantitative comparison between the efficiencies of the paper-based and CAAT-driven audit processes because the HIV dataset we audited on paper differed from the TB and Cancer datasets. The quantity and types of data elements varied among datasets, as did the types of source documents we reviewed to validate them. The auditors had also developed greater auditing skill and knowledge of the domain by the second round of audits, which would have biased the results in favor of the audit tool.

Conclusion

Audits are an important quality assessment and intervention, but paper audit forms complicate the audit process, during the preparation for and conduct of the audit, as well as the post-audit analysis. We identified five attributes that would allow a computer application to replace paper audit forms: computer networking for simultaneous application use by multiple auditors, standard formats for importing data and metadata, a structured approach to classifying data errors, support for pre-audit record selection, and the automatic tabulation of audit results. Our prototype CAAT conformed to these specifications and measurably improved the efficiency of our source document reviews. Furthermore,

use of an electronic system enforced consistency in our audit forms, error coding, and error tallying, which likely improved the comparability of our audit results across sites.

After extending the user interface to handle different types of data and completing a cycle of code refactoring, we intend to test the applicability of this simple, computer-based audit tool in networks outside of CCASAnet and on datasets of non-HIV data. These steps will allow us to evaluate the validity and comprehensiveness of our five CAAT desiderata and identify extensions to the application that increase its potential for general use. Our experiment demonstrated that a computerized audit tool can simplify the audit process and enable research networks to measure and improve the quality of their data. We hope other similar networks will find a computer-based auditing approach useful for evaluating the quality of their data, standardizing their quality control activities, and identifying areas for process improvement.

CHAPTER VI

EVALUATING DATA QUALITY DIMENSIONS IN THE CONTEXT OF CLINICAL RESEARCH DATA AUDITS

Introduction

Comparing the quality of research data generated by different clinical sites requires a thorough and standardized method of classifying characteristics of the audited data, especially properties related to data inconsistency and error [53]. Unfortunately most error classification schemes are designed for a specific task and lack generalizability [23]. The U.S. Department of Defense, for example, maintains a list of codes for labeling specific discrepancies in military inventory and accounting data [131], while the National Coordinating Council for Medication Error Reporting and Prevention offers a specialized taxonomy for tracking medication errors in healthcare organizations [132]. Other healthcare error taxonomies focus on medical device errors, patient safety, adverse events, and quality of care [133-136]. Information quality researchers have derived a detailed "taxonomy of dirty data" using theoretical principles and constructed informal taxonomies based on expert opinion, but their classification systems model errors within enterprise data warehouses and not discrepancies between medical research data and their physical source documents [137,138].

Clinical researchers often judge the quality of study data by their accuracy and completeness. Such aspects of quality that require use of different techniques to measure or document them are often referred to as dimensions of data quality [139]; over 26 different dimensions are noted in the scientific literature, with accuracy and completeness among the most cited [20]. Kahn et al., for example, defined twelve dimensions of quality relevant to information stored in product and consumer databases:

accessibility, appropriate amount of information, believability, completeness, concise representation, consistent representation, ease of manipulation, free-of-error, interpretability, objectivity, relevance, reputation, security, timeliness, and "value-added" [24]. Redundant naming schemes are rampant, however; studies by Wang, Kriebel, and Kahn use the terms accuracy, correctness, and "free-of-error" respectively to refer to truthful and accurate data [24,140].

Data quality is a context-dependent concept and what comprises data quality-related information for clinical source document verification audits -- beyond assessments of accuracy and completeness -- is not well documented in the literature [141]. We believe understanding data quality dimensions relevant to clinical research data auditing can facilitate more effective quality control processes. Auditors can design better audit methodologies, prioritize evaluation of the most relevant aspects of data quality, and capture quality-related information in a consistent, standardized format. Strong and Wang have proposed a framework for identifying aspects of data quality that are relevant to end users by examining organizations' quality assessment projects and using qualitative analysis to identify patterns of data quality problems [141].

We explore the application of this framework to identify a set of quality features relevant to clinical research data auditing. This study examines the quality assessment projects conducted by the CCASAnet coordinating center and attempts to determine what quality attributes auditors consider when determining whether data are fit for use.

Study Subject: CCASAnet Audit Datasets

The CCASAnet project aims to create a shared repository of HIV data from participating sites in Latin America and the Caribbean and to use the combined data to "answer questions about the characteristics of the regional HIV epidemic" [82]. Current CCASAnet membership consists of HIV/AIDS

centers at seven hospitals and clinics in Argentina, Brazil, Chile, Haiti, Honduras, Mexico, and Peru, plus a data coordinating center (DCC) at Vanderbilt University in Nashville, USA. A team from Vanderbilt conducted on-location, source verification data audits of datasets submitted to the DCC in response to proposed CCASAnet projects.

The CCASAnet DCC audited four distinct datasets, labeled HIV, HIV-study, Cancer, and Tuberculosis (TB) between 2007 and 2010. The HIV dataset submitted by each participating CCASAnet clinic included all HIV-positive, treatment-naïve patients beginning HAART, as well as an unlimited number of data points for demographics, treatment history, opportunistic infections, and laboratory test results. The site determined the contents and format of the variables, usually based on the structure of the local research database. The second dataset, HIV-study, contained excerpts from the HIV dataset that were reformatted by the sites and the DCC for use in two core CCASAnet research projects. Site personnel constructed the remaining two datasets, Cancer and TB, for two CCASAnet-specific projects on malignancy and tuberculosis in HIV-infected individuals. All data were numeric, categorical, and text data, rather than multimedia (e.g., images, video).

The DCC provided structured case report forms for the cancer and TB projects and site personnel submitted the data using REDCap's web-based data collection forms. The DCC conducted all HIV, Cancer, and TB data audits; site personnel and the DCC audit team independently audited the HIV-study datasets.

For all paper-based audits, auditors noted data errors and quality concerns by recording a corrected value and any necessary descriptive comments. Error codes, listed in Table 10, were applied to the data post-audit. For all computer-based audits, auditors coded errors on-site and recorded corrected values and comments using the CAAT web forms. Auditors could determine a particular value

was in error (using a specific error code) and optionally add quality-related comments. Values that were labeled as correct (code 1) could also generate quality-related comments.

Table 10: EORTC audit codes
Numeric audit codes from the European Organization for the Research and Treatment of Cancer, and code definitions adapted for audits of clinical research databases.

Audit Code	EORTC Definitions (adapted for an audit of database content)
1	*correct*: data in the database fully corresponds with the information in the medical file
2	*minor error*: database value does not correspond with information in the medical file, but the discrepancy is of minor clinical significance
3	*major error*: database value does not correspond with information in the medical file and the discrepancy is clinically significant
4	*missing/missed*: requested data are missing in the database
5	*sourceless*: data are present in the database, but not in the medical record (impossibility of checking correctness of the data)

Methods

We conducted a retrospective review of all audit forms completed during three years of CCASAnet data audits at member sites in Argentina, Brazil, Chile, Haiti, Honduras, Mexico, and Peru. For the original HIV and HIV-study audits we reviewed paper forms; for audits of the Cancer and TB datasets, we reviewed the data entry screens and printouts from the computer-assisted audit tool described in Chapter V. We inspected data types and values, error classifications, and auditor notes, both notes specific to individual data points as well as free text comments about each record as a whole. Variables included demographics, clinical data (e.g., weight, HIV staging, clinic visit dates), laboratory tests (e.g., CD4 lymphocyte count, HIV viral load, hemoglobin, TB culture, or biopsy results), ARV or TB medication regimens, cancer treatments, and all associated dates.

For each record, we noted whether the error ratings and auditor notes suggested information on data quality that was not captured in a structured or standardized fashion by the audit process. We

counted recurrences of similar auditor observations and grouped them by themes using a bottom-up approach. Finally, we attempted to map our quality themes to the 26 most frequently cited dimensions of data quality, as described in the data quality dimension literature reviews of Wand and Knight [20,142].

Results

Our manual review of audit documentation covered 520 audit forms that had been completed during 25 on-site audits by four auditors from the DCC and five internal auditors at CCASAnet sites A, D, E, and G. Of the audit documents, 378 were paper forms and 162 were printouts from the computer-based audit tool (Table 11). These forms contained audit findings for approximately 17,000 variables.

Table 11: Sources and types of audit documentation included in data quality review
*Totals are not additive due to overlap; CCASAnet comprises only seven sites and the same DCC audit team participated in every dataset audit.

Dataset	Audit Sites	Auditors	Forms Reviewed	Format
Cancer	6	3	72	electronic
TB	6	3	90	electronic
HIV	7	3	223	paper
HIV-study	4	8	135	paper
Total	**7***	**9***	**520**	

We identified 597 auditor comments or markings in the audit documents that reflected quality issues not captured by our error classification system. Not all quality comments were attached to data errors, and not all data errors had auditor comments. These observations represented 18 categories of quality comments and included auditor notes about the quality and type of source documents, date estimations, duplicate values, and predictions of the error source.

We mapped 17 categories to eight different dimensions of data quality: *reputation* (the degree of respect for the data's quality based on the information source), *precision* (the degree of exactness with which a value is stated), *consistency* (the degree to which data are "presented in the same format and compatible with previous data"), *relevance* (the degree to which data are "applicable and helpful" for the study), *timeliness* (the degree to which the data are "sufficiently up-to-date" for the study), *objectivity* (the degree to which data are "unbiased, unprejudiced, and impartial"), *conciseness* (the degree to which the data are "compactly represented without being overwhelming"), and *clarity* (the degree to which data are "easily comprehended" and understandable without ambiguity) [140,142]. Descriptions of the auditor observations we identified, their related dimensions of data quality, and the number of observations in each category are shown in Table 12.

Of the auditor observations, 173 (29%) reflected confidence in or concern for the quality and types of source documents against which the data were being validated. The majority of these comments indicated what type of source document was used to confirm a data point ("mentioned in visit note" or "confirmed by laboratory database") or qualified an auditor's "no source" declaration (code 5) by noting whether he'd conducted a cursory or an exhaustive search before applying the label. Other auditors used margin notes to remark on temporarily unavailable source documents. One site, for example, could only provide the audit team with clinic charts and not the corresponding hospital charts because employees in the hospital records office were on strike. The auditors noted the problem but chose not to label affected data with error codes because it would inflate the site's true "no source" rate and diminish the usefulness of audit findings.

Table 12: Types of data quality observations not captured by the standard audit protocol, their frequency, and related data quality dimensions

This table lists the 18 categories of quality-related observations we found in our review CCASAnet audit documents as well as their description, frequency, percentage of 597 total comments, and related dimension of data quality.

Dimension	Observation	*N* (%)	Description of quality comment
Reputation	Quality or Type of Source Document	92 (15%)	Notes mention what type of source document provides proof of data validity or error (e.g., lab report, flowsheet, clinic intake sheet, clinic note, etc.).
	Partial Sources	29 (4.9%)	Only partial source documents are available (e.g., clinic chart but not hospital chart).
	Auditor Search Intensity	52 (8.7%)	"No source" (code 5) assignments include a statement of how thoroughly the auditor searched for a source document in the CR.
Precision	Rounded Values	65 (11%)	Values are rounded although exact measurements are available.
	Approximations	45 (7.5%)	DB contains approximate values even though exact ones are available in the CR, or approximate values aren't labeled as such.
	No Dates	19 (3.2%)	Lab values, patient ages, or other data points that require a date to have meaning have no associated date.
	Overprecision	27 (4.5%)	Data are reported with false exactness
	Preferable Value	18 (3.0%)	Other lab values are better choices given the specified timeframe.
Consistency	Inconsistent Definitions	43 (7.2%)	The definitions for this value vary depending on who recorded the information.
	Inconsistent Formats	38 (6.4%)	Data are recorded using inconsistent codes or syntax (e.g., alternating date formats, male/masculine/M)
	Conflicting Data	27 (4.5%)	The dataset lacked internal consistency; some values contradicted other values.
Relevance	Inclusion of Record	21 (3.5%)	The record does not meet the study's inclusion criteria and should not be present in the dataset.
	Irrelevant Data	4 (0.6%)	The record contains unnecessary information that is not required for any study and was not requested by the DCC.
Timeliness	Outdated Data	24 (4.0%)	The record hasn't been updated in many years despite new information, or the record isn't as current as suggested by the "last update date" variable.
Objectivity	Interpreted Data	14 (2.3%)	The variable contains interpreted data rather than raw data. (i.e., Instead of a numeric lab value, field says "normal").
Conciseness	Duplicates	3 (0.5%)	The dataset included multiple, redundant instances of a single value.
Clarity	Vague Data	8 (1%)	The meaning of a value was ambiguous without extra data.
N/A	Prediction of Error Cause	68 (11%)	Miscellaneous auditor comments speculating on potential causes of error (e.g., typographical slip, distorted rubber date stamp)
Total	18 types	597 (100%)	

An additional 174 (29%) audit comments noted imprecise representations of data, including 65 cases of weight, creatinine, and hemoglobin measurements that had been rounded to the nearest integer (and sometimes *not* the nearest integer, as in a case of 76.8 kg rounded down to 76 kg.) Other instances of imprecision occurred when site personnel submitted medication regimens with approximated start and stop dates despite having exact dates documented in on-site pharmacy records. The case report forms for the Cancer and TB data collections requested laboratory values "at baseline," defined as a 30-day window preceding HIV or TB treatment start, and site abstractors occasionally submitted valid but non-optimal data selections for these data. In one case, the record abstractor chose a viral load result taken 28 days before baseline over a value from 2 days before baseline. While the submitted, 28-day value was not incorrect, the 2-day result more precisely represented "viral load at baseline."

We also noted 27 observations of false precision in the audit data. Two sites frequently failed to identify estimated dates despite a checkbox on the data collection form indicating "this date is an approximation." The resulting data points falsely appeared to be exact dates. A third site's data submission contained CD4 cell counts and hemoglobin measurements with two decimal place precision even though the laboratory reported only integers; a CD4 value of 98 cells/mm^3 was stored as 98.44 in the database. (The site's data manager later established that the data distortions were due to faulty coding in the lab-to-database import script.)

The third most common category of auditor observations concerned the consistency of data encountered during the audits, particularly in data definitions, data representation, and dataset content (108 auditor comments, 18%). In 43 cases, clinicians recorded assessments of hepatitis B infection or HIV staging using definitions that did not correspond with the majority of the dataset's records. In one site's dataset, most "patient HIV stage" variables were coded using World Health Organization criteria, but a

few cases used criteria published by the U.S. Centers for Disease Control and Prevention instead [143,144]. Two medical record abstractors from a different CCASAnet clinic correctly interpreted the “Site & Subsite” prompt of the Cancer report form as referring to the part of the body where the cancer was located. The third abstractor mistook the prompt (“Site”) for a question about the hospital where the cancer was diagnosed.

Additionally, 38 records from two datasets in our review recorded semantically identical data content with alternating, inconsistent syntax, using character strings like “M,” “male,” or “masculino” to represent the same “patient gender” concept, or “<80,” “79,” or “0” to code an undetectable viral load. Other auditors noted contradictory values in audit material, such as visit dates after the patient’s death or opposing variables such as “HAART naïve” and “previous HAART” both coded as “yes.”

Auditor observations concerning the relevance or timeliness of data each occupied four percent of the data quality comments. In 25 cases, records failed the inclusion criteria for their respective datasets and data were irrelevant to the proposed studies. In another 24 cases, the patient information submitted to the DCC was extremely outdated and not as recent as the record’s “last update date” variable suggested. One patient had died six months prior to a planned record update, but two years later his death was still not reflected in his electronic file.

Physicians’ interpretations compromised data objectivity in 14 cases; affected variables contained interpreted data rather than raw data. Lab results, for example, were recorded as "normal" rather than numeric values. Our retrospective review also detected three cases in which auditors remarked upon a lack of concise data representation, particularly with duplicate data points. In 8 remaining quality-related observations, auditors copied additional information from the clinical record to compensate for lack of clarity in the audit data.

We also identified 68 records that contained auditor notes concerning potential causes for observed errors. These included typographical mistakes in case report form-to-database transcription, incorrectly sorted lab slips, or misshapen rubber date stamps that inked illegible numbers.

Discussion

Summary of Findings

Our analysis identified eight dimensions of data quality that were ineffectively captured using either the traditional paper audit or the computer assisted audit tool, namely the clarity, timeliness, concise representation, consistent representation, objectivity, relevance, precision, and reputation of the data. Almost a third of the audit observations concerned the trustworthiness of the data relative to the quality of the source document. The auditors' comments reflected awareness that an original lab slip confers greater trust in the accuracy and authenticity of a laboratory result than a handwritten doctor's note in the patient record. However, the "no source" (code 5) error classification was only a dichotomous measure of data reputation. Auditors also hesitated to label a variable with a code 5 error if they had not exhaustively searched the clinical record for a potential source document. The quantity of such observations suggests auditors require a more granular measure of data reputation, both to code quality observations and manage auditors' time, as data elements with low relevance to study outcomes may not necessitate an exhaustive search for documentation.

Other deficiencies of the datasets, such as inconsistent variable representations, outdated content, and imprecise data could delay or bias the results of research studies: when such errors are apparent, study progress is delayed until research sites provide the coordinating center with data

corrections; alternately, the study could be compromised by poor quality data and investigators might remain unaware until subsequent data quality inspections.

Limitations

Our study investigated the dimensions of data quality relevant to source document verification audits in the context of an observational research network for HIV epidemiology composed of member sites in resource-limited settings using paper-based medical records. Without additional investigation, we cannot know whether the aspects of data quality we identified generalize to source verification audits of other datasets in other settings. Indeed, analyzing audits of clinical sites whose primary source documents are electronic systems may reveal additional dimensions of data quality that affect whether the resultant datasets are fit for use.

One researcher identified all quality-related comments in the audit documents and assigned them to classes of "audit observations" and "data quality dimensions." Validation by a second analyst would increase the legitimacy of these results.

This retrospective documentation review likely underestimated the variety of quality dimensions in clinical research data audits. We are certain auditors did not document every data quality-related comment, but they are likely to have recorded the most critical or frequently occurring quality observations. Our findings, therefore, may represent a critical but not exhaustive list of data quality dimensions relevant to source verification auditing. Collecting feedback from auditors during the audit process might increase the numbers of observations and reveal additional categories of error.

Conclusion

Our study established that auditors consider dimensions in addition to accuracy and completeness when determining whether study data are fit for use. Clinical research data auditing involves at least eight additional data quality dimensions, including reputation, precision, consistency, relevance, timeliness, objectivity, conciseness, and clarity. The EORTC error coding system includes only dichotomous markers of accuracy, completeness, and "existence of source document," and therefore cannot represent the complexity of these additional dimensions. This limitation leads auditors to record their observations as free-text notes.

By capturing additional dimensions of data quality formally and routinely, auditors can better characterize quality variances among sites and offer suitable recommendations. When observing inconsistencies in data definitions or formatting, for example, auditors can suggest the affected site develop a handbook of standard operating procedures for data recording and capture. Since regular assessment of all relevant dimensions could burden auditors with memorizing lists of data error types and frequent quality observations, prompts and tools to assess such dimensions should be incorporated in an auditing CAAT. The error coding form of the system described in Chapter V might include a dropdown selection for "source type/quality", a slider bar for "intensity of auditor search", or a user tagging system to mark unnecessary approximations or probable typographical slips.

Source verification audits like those conducted by CCASAnet could render comprehensive and informative results more efficiently if relevant data quality dimensions were captured in a standardized fashion. Auditors could benefit from a deeper understanding of what aspects of quality render a dataset fit for use. A comprehensive quality assessment framework for clinical research data auditing must be flexible and enable auditors to measure and record many observations on data quality beyond simple measures of accuracy and completeness.

CHAPTER VII

DISCUSSION

Summary of Findings

Patient records maintained by healthcare providers offer an abundance of longitudinal medical data that potentially can be reused for research. Such data fuel the studies conducted by many large, global consortia, including major regional cohorts that have been established to address treatment and outcomes in HIV-positive persons. The process of transforming patchwork hospital and clinic documentation into uniform, formatted research data is challenging, particularly in resource-limited settings where data are often of poor quality. Nevertheless, investigators rarely conduct systematic data quality assessments when the record abstraction occurs outside the strict monitoring programs of clinical trials or when pre-existing databases are recycled for new studies.

We conducted a series of source verification data audits to determine the accuracy and completeness of data submitted to CCASAnet, an international, multicenter observational research network for HIV epidemiology. Our analysis identified rates of missing weight measurements, unverifiable laboratory values, and discrepancies in antiretroviral treatment regimens and dates that would have triggered quality violations and immediate intervention in the context of controlled clinical trials. Database integrity checks performed at the DCC would not have identified most of these errors, since the data appeared normal and internally consistent. These findings illustrated a pressing need for *on-site* data quality assurance activities in networks like CCASAnet.

Local research teams reported on potential sources of flawed data in a post-audit survey, naming difficulties with the syntax and semantics of the data and the culture of the workplace. Although

the direct causes of missing source documents or data were often identical across sites, the root causes were site-specific, resulting from limited clinic resources, workflow complications, or the hostility of external participants towards research activities. The sources of error identified by site personnel informed future predictions of error causes documented by the DCC audit team.

Our initial audit experiences underscored the limitations of paper audit forms for standardizing and reporting audit findings. We analyzed the CCASAnet audit process to identify key attributes for a computer-assisted audit tool, which included networking, audit data management, standardized error assessment, decision support, and results reporting tools. We developed a prototype application based on these desiderata and piloted it during six audits of tuberculosis and cancer data from CCASAnet clinics. Our informal assessments suggested the CAAT improved the efficiency, clarity, and completeness of our data audits.

Finally, we cataloged issues in data quality that were not adequately represented by standard measures of accuracy and completeness. We discovered auditors made margin notes related to eight additional dimensions of data quality, especially the precision of data, consistency of representation, and reputation of the source document. These findings clarified what makes CCASAnet data "fit for use" and suggested additional functionality for our computer-based audit application.

Study Limitations

Throughout this work we have treated the patient medical record as a gold standard for data quality, which is standard research data practice [55]. Medical records, however, are not complete histories of patient care due to differences in how individual doctors collect patient histories, conduct physical exams, interpret lab results, and document these activities [145]. A study by Rethans et al. found the correlation between the number of actions physicians took during patient consultations and

the number of actions documented in subsequent clinical notes was only 0.54, while the documentation provided insufficient information to measure clinician performance in clinical audits [146]. Abstraction of such incomplete medical records may introduce more errors [147]. Since medical records portray an incomplete picture of patient care, our audits may have underestimated the true rates of data variability and error in submitted datasets.

Even when using a standardized error classification system, as we did with the 5-stage quality codes from the European Organization for the Research and Treatment of Cancer (EORTC), different approaches to labeling errors and counting the total number of data fields (for the denominator of the error rate) can result in notably different error rates [49]. The error rates reported in Chapter II, the questions asked of each site in the Chapter III survey, and the frequency of auditor observations documented in Chapter VI could be affected. We have attempted to minimize variations in the application of EORTC codes by making one person responsible for the initial coding of all errors reported here, with other auditors reviewing the work to confirm consistency and correctness.

The limited perspective of this research also affects its external validity; CCASAnet was the setting for all audits and all datasets were related to the care of HIV-positive persons. Audit experiences with other collaborations and datasets might reveal additional categories of data error, desiderata for an audit tool, and relevant dimensions of data quality.

Although we have confidence in the effectiveness of auditing, we cannot demonstrate that our quality assessment procedures have had a measurable impact on data quality at CCASAnet sites as we did not study the impact of reporting our audit findings or implementing local quality interventions. Although our eventual goal is to improve provider documentation and research data procurement measurably at CCASAnet sites, simply providing investigators with estimates of the error rates in their study data can have a positive impact on their data selection and study design [148].

Study Implications and Future Work

This project represents innovative work in the field of audit informatics. The collection of studies presented here comprises an initial effort to detail data quality problems in HIV observational datasets and to describe, simplify, and standardize the audit process for clinical research data. Our findings should encourage observational networks to establish central data monitoring programs and guide clinic sites in conducting routine quality assessment activities.

Future work will seek to quantify the costs of auditing in observational research networks, while exploring innovative low-cost approaches. We also intend to develop and distribute an open source software application for data quality auditing and formally evaluate its effectiveness in comparison to paper-based audits. Our prototype application proved effective and the work described in Chapters IV and VI provided additional auditing insights to guide the development of a comprehensive CAAT. Sites without a trained data manager could benefit from using such an application to establish an internal quality assessment program. By observing diverse applications of the software, we can assess whether the audit tasks and relevant aspects of data quality we have described generalize to settings beyond CCASAnet.

Verifying data against source documents through audits will improve the quality of databases and research and can be a technique for retraining staff responsible for clinical data collection. Audit efforts may be optimized by using software that supports key audit functions and allows documentation of diverse data errors. We recommend that all participants in observational cohorts use computer-based data audits to assess and improve the quality of data and to guide future data collection and abstraction efforts at the point of care.

APPENDIX A

EXAMPLE OF AUDIT FORM AND ERROR CODING

Table 13: Example of the Demographics and Laboratory Data sections of a completed audit form
This sample form demonstrates how errors in clinical and laboratory data are coded and documented when auditors identify values in clinical records that conflict with those recorded in the database.

Test Record #1	Value in database	Value in clinical record	Comments	Audit Code
Demographics and clinical data				
Gender	Male	Male		correct (1)
Birthdate	1973-01-31	1973-01-31		correct (1)
Weight	56	56.5	Rounding	minor error (2)
Weight date	24 Aug 2002	24 Aug 2002		correct (1)
Laboratory data				
CD4	---	110		missing (4)
CD4 date	---	15 Aug 2002		missing (4)
Viral load	32,000	320,000	Probable typo	incorrect (3)
Viral load date	15 Aug 2002	15 Aug 2002		correct (1)

Table 14: Example of the Antiretroviral Regimens sections of a completed audit form
This sample form demonstrates how ARV errors are coded and documented when auditors identify values in clinical records that conflict with those recorded in the database.

Test Record #1	Value in database	Value in clinical record	Comments	Audit Code
Regimen 1	AZT DDI RIT SAQ	AZT DDI RIT SAQ		correct (1)
Start date	28 Aug 2002	28 Aug 2002		correct (1)
Stop date	09 May 2006	06 Feb 2005	Notes from the HIV clinic show that regimen 1 was stopped on 06 Feb 2005. Only the Tuberculosis clinic continues to report regimen 1 active until 09 May 2006.	incorrect (3)
Regimen 2	---	AZT 3TC EFV		missing (4)
Start date	---	06 Feb 2005	(documented in hospital orders and notes from HIV clinic)	missing (4)
Stop date	---	01 Jan 2006		missing (4)
Regimen 3	---	D4T 3TC EFV KAL		missing (4)
Start date	---	14 Feb 2006	Regimen 3 began during hospitalization and is listed as dispensed on hospital orders checksheet	missing (4)
Stop date	---	21 Feb 2006	Regimen 3 stops appearing in hospital orders.	missing (4)
Regimen 4	D4T 3TC KAL	D4T 3TC KAL		correct (1)
Start date	09 May 2006	22 Feb 2006	Regimen 4 was begun on 22 Feb 2006 during hospitalization and (according to the order sheets) the drugs were administered.	incorrect (3)
Stop date	19 Sept 2006	23 Jul 2006	Clinic notes show Regimen 4 was stopped on 23 Jul 2006.	incorrect (3)
Regimen 5	AZT 3TC KAL	AZT 3TC KAL		correct (1)
Start date	19 Sep 2006	18 Sep 2006	Unclear handwriting.	minor error (2)
Stop date	current/ongoing	11 Dec 2007	Patient died.	incorrect (3)

APPENDIX B

EXAMPLES OF CAAT METADATA AND DATA SPECIFICATIONS

Table 15: Example of CAAT metadata specifications
This table presents an example of the data dictionary format imported by the computer-assisted audit tool (CAAT) described in Chapter V. Each row describes how to display one element on the web form shown in Figure 13 (see page 62).These specifications are based on the REDCap metadata format.

Variable / Field Name	Form Name	Section Header	Field Type	Field Label	Choices	Field Note
study_id	demographics		text	Study ID		
person_identifier	demographics	TB Case Abstract Form	text	ID number		
dob	demographics		text	Date of birth		MM/DD/YYYY or YYYY-MM-DD
dob_estimate	demographics		advcheckbox	Is the date above an estimate?	0, Unchecked | 1, Checked	
sex	demographics		dropdown	Sex	0, Male | 1, Female | 2, Transsexual	
bcg_history	demographics		dropdown	History of BCG vaccination	0, Yes | 1, No | 2, Unknown	
bcg_scar	demographics		dropdown	BCG scar	0, Yes | 1, No | 2, Unknown	
transmission	demographics		dropdown	Risk factor for HIV	0, Heterosexual | 1, Homosexual | 2, Bisexual | 3, IDU | 4, MTCT | 5, Other	
transmission_other	demographics		text	Other: (describe here)		
usual_occupation	demographics		text	Patient's usual occupation		

Table 16: Example of CAAT data specification
This table presents an example of data formatted for import into the computer-assisted audit tool. Categorical variables are coded according to lists of choices defined by the metadata format shown in Table 15.

study_ id	person_ identifier	dob	dob_ estimate	sex	bcg_ history	bcg_ scar	transmission	transmission_ other	usual_ occupation
1	test1	12/28/1981	0	0	1	1	1		taxi driver
2	test2	4/7/1974	0	0	2	2	3		
3	test3	2/15/1954	1	0	2	2	1		
4	test4	1/13/1949	0	1	2	2	0		sex worker
5	test5	11/23/1971	0	1	2	2	0		
6	test6	11/2/1981	0	0	2	2	2		stylist

REFERENCES

1. International Conference on Harmonisation. Harmonized Tripartite Guideline: Guideline for Good Clinical Practice [Internet]. 1996. Available from: http://www.ich.org/cache/compo/276-254-1.html

2. de Lusignan S, van Weel C. The use of routinely collected computer data for research in primary care: opportunities and challenges. Fam. Pract. 2006 Apr 1;23(2):253-263.

3. National Institutes of Health. Clinical Research Networks and NECTAR - Overview [Internet]. 2008 Jan 18 [cited 2008 May 15];Available from: http://nihroadmap.nih.gov/clinicalresearch/overview-networks.asp

4. Zerhouni E. MEDICINE: The NIH Roadmap. Science. 2003 Oct 3;302(5642):63-72.

5. Kuperman GJ, Blair JS, Franck RA, Devaraj S, Low AFH, for the NHIN Trial Implementations Core Services Content Working Group. Developing data content specifications for the Nationwide Health Information Network Trial Implementations. Journal of the American Medical Informatics Association. 2010 Jan 1;17(1):6 -12.

6. Luukkonen T, Tijssen RJW, Persson O, Sivertsen G. The measurement of international scientific collaboration. Scientometrics. 1993 9;28(1):15-36.

7. Wagner CS, Leydesdorff L. Network structure, self-organization, and the growth of international collaboration in science. Research Policy. 2005 Dec;34(10):1608-1618.

8. Leydesdorff L, Wagner CS. International collaboration in science and the formation of a core group. Journal of Informetrics. 2008 Oct;2(4):317-325.

9. Wagner CS, Leydesdorff L. Mapping the network of global science: comparing international co-authorships from 1990 to 2000. International Journal of Technology and Globalisation. 2005;1(2):185 - 208.

10. McKee M. Routine data: a resource for clinical audit? Qual Health Care. 1993 Jun;2(2):104-111.

11. The Antiretroviral Therapy in Lower Income Countries (ART-LINC) Study Group. Cohort Profile: Antiretroviral Therapy in Lower Income Countries (ART-LINC): international collaboration of treatment cohorts. International Journal of Epidemiology. 2005 Oct;34(5):979 -986.

12. Zhou J, Kumarasamy N, Ditangco R, Kamarulzaman A, Lee CKC, Li PCK, et al. The TREAT Asia HIV Observational Database: baseline and retrospective data. J Acquir Immune Defic Syndr. 2005 Feb 1;38(2):174-9.

13. Weiner MG, Embi PJ. Toward Reuse of Clinical Data for Research and Quality Improvement: The End of the Beginning? Annals of Internal Medicine. 2009;151(5):359 -360.

14. Beretta L, Aldrovandi V, Grandi E, Citerio G, Stocchetti N. Improving the quality of data entry in a low-budget head injury database. Acta Neurochir (Wien). 2007;149(9):903-909.

15. Curado M, Edwards B, Shin H, Storm H, Ferlay J, Heanue M, et al. Cancer Incidence in Five Continents Vol. IX [Internet]. 2007 [cited 2011 Feb 14]. Available from: http://www.iarc.fr/en/publications/pdfs-online/epi/sp160/index.php

16. U.S. Department of Health & Human Services. HHS Guidelines for Ensuring and Maximizing the Quality, Objectivity, Utility, and Integrity of Information Disseminated to the Public [Internet]. Washington, D.C.: 2002. Available from: http://aspe.hhs.gov/infoquality/Guidelines/index.shtml

17. Davis JR, Nolan VP, Woodcock J, Estabrook RW. Assuring Data Quality and Validity in Clinical Trials for Regulatory Decision Making:Workshop Report [Internet]. Washington, D.C.: The National Academies Press; 1999 [cited 2011 Feb 9]. Available from: http://www.nap.edu/catalog.php?record_id=9623

18. Grantner E. ISO 8000 - A Standard for Data Quality. Logistics Spectrum [Internet]. 2007 Oct 1;(Oct-Dec 2007). Available from: http://findarticles.com/p/articles/mi_qa3766/is_200710/ai_n27997242/?tag=content;col1

19. ISO - International Organization for Standardization, ISO - International Organization for Standardization. ISO/TS 800 - Data Quality -- Part 100: Master data: Overview [Internet]. [cited 2011 Feb 17];Available from: http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=52129

20. Wand Y, Wang RY. Anchoring data quality dimensions in ontological foundations. Commun. ACM. 1996;39(11):86-95.

21. Tupek AR, Census Bureau Methodology & Standards Council. Census Bureau Principle: Definition of Data Quality. 2006.

22. Arts DGT, de Keizer NF, Scheffer G. Defining and Improving Data Quality in Medical Registries: A Literature Review, Case Study, and Generic Framework. Journal of the American Medical Informatics Association. 2002 Nov 1;9::600 -611.

23. Pipino LL, Lee YW, Wang RY. Data quality assessment. Commun. ACM. 2002;45(4):211-218.

24. Kahn BK, Strong DM, Wang RY. Information quality benchmarks: product and service performance. Communications of the ACM. 2002 Apr;45:184–192.

25. Lee YW, Strong DM, Kahn BK, Wang RY. AIMQ: a methodology for information quality assessment. Inf. Manage. 2002;40(2):133-146.

26. Chapman AD, Global Biodiversity Information Facility. Principles of data quality. Copenhagen :: Global Biodiversity Information Facility; 2005.

27. Duran E. Customer Data Management: The Impacts of Poor Customer Data Quality on Customer Relationship Management and Organizations. Saarbrücken, Deutschland: VDM Verlag Dr. Müller; 2009.

28. Malcom A. Poor data quality costs 10% of revenues, survey reveals [Internet]. Computerworld. 1998 Jul 20 [cited 2011 Feb 15];Available from: http://computerworld.co.nz/news.nsf/UNID/CC256CED0016AD1ECC25684C000E0278?OpenDocument&Highlight=2,Poor,data,quality,costs

29. Wolff H, Chong H, Auffhammer M. Classification, Detection and Consequences of Data Error: Evidence from the Human Development Index. National Bureau of Economic Research Working Paper Series [Internet]. 2010 Dec [cited 2011 Jan 27];No. 16572. Available from: http://www.nber.org/papers/w16572

30. Berner ES, Kasiraman RK, Yu F, Ray MN, Houston TK. Data Quality in the Outpatient Setting: Impact on Clinical Decision Support Systems. AMIA Annu Symp Proc. 2005;2005:41-45.

31. Weiss RB, Gill GG, Hudis CA. An On-Site Audit of the South African Trial of High-Dose Chemotherapy for Metastatic Breast Cancer and Associated Publications. J Clin Oncol. 2001 Jun 1;19(11):2771-2777.

32. Ancukiewicz M, Niemierko A, Goldberg S. The Impact of Data Entry Errors on the Findings in Radiation Oncology Research. International Journal of Radiation Oncology*Biology*Physics. 2005 Oct 1;63(Supplement 1):S191-S192.

33. Mullooly JP. The effects of data entry error: An analysis of partial verification. Computers and Biomedical Research. 1990 Jun;23(3):259-267.

34. Chang YC, Kim JD, Schwander K, Rao DC, Miller MB, Weder AB, et al. The impact of data quality on the identification of complex disease genes: experience from the Family Blood Pressure Program. Eur J Hum Genet. 2006 Feb 22;14(4):469-477.

35. Abecasis GR, Cherny SS, Cardon LR. The impact of genotyping error on family-based analysis of quantitative traits. Eur. J. Hum. Genet. 2001 Feb;9(2):130-134.

36. Larsen IK, Småstuen M, Johannesen TB, Langmark F, Parkin DM, Bray F, et al. Data quality at the Cancer Registry of Norway: An overview of comparability, completeness, validity and timeliness. European Journal of Cancer. 2009 May;45(7):1218-1231.

37. Brewster D, Crichton J, Muir C. How accurate are Scottish cancer registration data? Br J Cancer. 1994 Nov;70(5):954-959.

38. Vantongelen K, Rotmensz N, Van Der Schueren E. Quality control of validity of data collected in clinical trials. European Journal of Cancer and Clinical Oncology. 1989 Aug;25(8):1241-1247.

39. Christian MC, McCabe MS, Korn EL, Abrams JS, Kaplan RS, Friedman MA. The National Cancer Institute Audit of the National Surgical Adjuvant Breast and Bowel Project Protocol B-06. N Engl J Med. 1995 Nov 30;333(22):1469-1475.

40. Forster M, Bailey C, Brinkhof MW, Graber C, Boulle A, Spohr M, et al. Electronic medical record systems, data quality and loss to follow-up: survey of antiretroviral therapy programmes in resource-limited settings. Bull World Health Organ. 2008 Dec;86(12):939-947.

41. Prins H, Kruisinga F, Bueller HA, Zwetsloot-Schonk JHM. Availability and usability of data for medical practice assessment. Int J Qual Health Care. 2002 Apr 1;14(2):127-137.

42. Hogan WR, Wagner MM. Accuracy of Data in Computer-based Patient Records. J Am Med Inform Assoc. 1997;4(5):342-355.

43. Pringle M, Ward P, Chilvers C. Assessment of the completeness and accuracy of computer medical records in four practices committed to recording data on computer. The British Journal of General Practice. 1995 Oct;45(399).

44. Jensen AR, Overgaard J, Storm HH. Validity of breast cancer in the Danish Cancer Registry. A study based on clinical records from one county in Denmark. Eur. J. Cancer Prev. 2002 Aug;11(4):359-364.

45. Nahm ML, Pieper CF, Cunningham MM. Quantifying data quality for clinical trials using electronic data capture. PLoS ONE. 2008;3(8):e3049.

46. Demlo LK, Campbell PM, Brown SS. Reliability of Information Abstracted from Patients' Medical Records. Medical Care. 1978 Dec 1;16(12):995-1005.

47. Lorenzoni L, Da Cas R, Aparo U. The quality of abstracting medical information from the medical record: the impact of training programmes. Int J Qual Health Care. 1999 Jun 1;11(3):209-213.

48. Colin C, Ecochard R, Delahaye F, Landrivon G, Messy P, Morgon E, et al. Data Quality in a DRG-Based Information System. International Journal for Quality in Health Care. 1994;6(3):275 -280.

49. Rostami R, Nahm M, Pieper CF. What can we learn from a decade of database audits? The Duke Clinical Research Institute experience, 1997--2006. Clin Trials. 2009 Apr;6(2):141-150.

50. Damkier P. Chapter 2: Good Clinical Practice [Internet]. In: The IUPHAR Compendium of Basic Principles For Pharmacological Research in Humans. p. 5.Available from: http://www.iuphar.org/pdf/hum_11.pdf

51. George SL. Clinical Trials Audit and Quality Control [Internet]. In: Armitage P, Colton T, editors. Encyclopedia of Biostatistics. Chichester, UK: John Wiley & Sons, Ltd; 2005 [cited 2011 Feb 14]. Available from: zotero://attachment/11562/

52. Weiss RB, Vogelzang NJ, Peterson BA, Panasci LC, Carpenter JT, Gavigan M, et al. A successful system of scientific data audits for clinical trials. A report from the Cancer and Leukemia Group B. JAMA. 1993 Jul 28;270(4):459-464.

53. Califf RM, Karnash SL, Woodlief LH. Developing systems for cost-effective auditing of clinical trials. Controlled Clinical Trials. 1997 Dec;18(6):651-660.

54. Neaton JD, Duchene AG, Svendsen KH, Wentworth D. An examination of the efficiency of some quality assurance methods commonly employed in clinical trials. Statistics in Medicine. 1990;9(1-2):115-124.

55. McFadden E. Management of data in clinical trials. Wiley-Interscience; 2007.

56. Dana-Farber/Harvard Cancer Center. Clinical Trial Standards Applied During Audit Proceedings [Internet]. 2009 [cited 2011 Mar 11];Available from: http://www.dfhcc.harvard.edu/fileadmin/DFHCC_Admin/Clinical_Trials/QACT/Policies_and_Procedures/CT_Standards_Applied_During_Audit_Proceedings2005.pdf

57. Reisch LM, Fosse JS, Beverly K, Yu O, Barlow WE, Harris EL, et al. Training, Quality Assurance, and Assessment of Medical Record Abstraction in a Multisite Study. American Journal of Epidemiology. 2003 Mar 15;157(6):546 -551.

58. Whitney CW, Lind BK, Wahl PW. Quality assurance and quality control in longitudinal studies. Epidemiol Rev. 1998;20(1):71-80.

59. Van den Broeck J, Mackay M, Mpontshane N, Kany Kany Luabeya A, Chhagan M, Bennish ML. Maintaining data integrity in a rural clinical trial. Clinical Trials. 2007 Oct 1;4(5):572 -582.

60. Vantongelen K, Steward W, Blackledge G, Verweij J, Van Oosterom A. EORTC joint ventures in quality control: Treatment-related variables and data acquisition in chemotherapy trials. European Journal of Cancer and Clinical Oncology. 1991;27(2):201-207.

61. Dicksee LR, Montgomery RH. Auditing: A Practical Manual for Auditors. New York: 1905.

62. Hysong SJ. Meta-analysis: audit and feedback features impact effectiveness on care quality. Med Care. 2009 Mar;47(3):356-63.

63. Probitts J. Auditing in the manufacturing environment. The Quality Assurance Journal. 2000;4(4):193-196.

64. Dark M, Poftak A. How to Perform a Security Audit. Technology and Learning. 24(7):18-27.

65. Cimino JJ. Auditing the Unified Medical Language System with Semantic Methods. J Am Med Inform Assoc. 1998;5(1):41-51.

66. Mougin F, Bodenreider O. Auditing the NCI Thesaurus with Semantic Web Technologies. AMIA Annu Symp Proc. 2008;2008:500-504.

67. Grider D. Medical Record Auditor: Documentation rules and rationales with exercises. 2nd ed. USA: American Medical Association; 2008.

68. Johnston G, Crombie I, Alder E, Davies H, Millard A. Reviewing audit: barriers and facilitating factors for effective clinical audit. Quality in Health Care : QHC. 2000 Mar;9(1).

69. Research Program Social Aspects of HIV/AIDS and Health. Audit of HIV/AIDS Policies: In Botswana, Lesotho, Mozambique, South Africa, Swaziland, and Zimbabwe. Human Sciences Research Council; 2005.

70. Weiss RB. Systems of protocol review, quality assurance, and data audit. Cancer Chemotherapy and Pharmacology. 1998;42(0):S88-S92.

71. Williams CA, Mosley-Williams AD, Overhage JM. Arthritis Quality Indicators for the Veterans Administration: Implications for Electronic Data Collection, Storage Format, Quality Assessment, and Clinical Decision Support. AMIA Annu Symp Proc. 2007;2007:806-810.

72. U.S. Department of Veterans Affairs. Veterans Affairs Performance and Accountability Report. Part II: Assessment of Data Quality [Internet]. 2010 [cited 2011 Mar 3]. Available from: http://www.va.gov/BUDGET/docs/report/PartII/FY2010-VAPAR_PartII_Assessment_of_Data_Quality.pdf

73. Smith CJ, Olsen CH, Mocroft A, Viard JP, Staszewski S, Panos G, et al. The role of antiretroviral therapy in the incidence of pancreatitis in HIV-positive individuals in the EuroSIDA study. AIDS. 2008 1;22(1):47-56.

74. Abdel-Khalik AR, Solomon I. Research Opportunities in Auditing: The Second Decade. Sarasota, FL: Amer Accounting Assn; 1989.

75. Glick J. On the potential cost effectiveness of scientific audits. Accountability in Research: Policies and Quality Assurance. 1989;1:37-43.

76. de Lusignan S, Stephens PN, Adal N, Majeed A. Does Feedback Improve the Quality of Computerized Medical Records in Primary Care? J Am Med Inform Assoc. 2002;9(4):395-401.

77. Shamoo AE. Principles of research data audit. Taylor & Francis US; 1989.

78. Shamoo AE. Data audit would reduce unethical behaviour. Nature. 2006 Feb 16;439(7078):784.

79. Hearnshaw H, Harker R, Cheater F, Baker R, Grimshaw G. Are audits wasting resources by measuring the wrong things? A survey of methods used to select audit review criteria. Qual Saf Health Care. 2003 Feb;12(1):24-28.

80. Rennie D. Accountability, Audit, and Reverence for the Publication Process. JAMA: The Journal of the American Medical Association. 1993 Jul 28;270(4):495 -496.

81. Lyon L. Dealing with Data: Roles, Rights, Responsibilities and Relationships [Internet]. 2007 [cited 2011 Feb 6]. Available from: http://www.ukoln.ac.uk/ukoln/staff/e.j.lyon/reports/dealing_with_data_report-final.pdf

82. McGowan CC, Cahn P, Gotuzzo E, Padgett D, Pape JW, Wolff M, et al. Cohort Profile: Caribbean, Central and South America Network for HIV research (CCASAnet) collaboration within the

International Epidemiologic Databases to Evaluate AIDS (IeDEA) programme. Int. J. Epidemiol. 2007 Sep 10;:dym073.

83. Tuboi S, Schechter M, McGowan C, Cesar C, Krolewiecki A, Cahn P, et al. Mortality during the first year of potent antiretroviral therapy in HIV-1-infected patients in 7 sites throughout Latin America and the Caribbean. J Acquir Immune Defic Syndr. 2009;51(5):615-623.

84. Cesar C, Shepherd B, Krolewiecki A, Fink V, Schechter M, Tuboi S, et al. Rates and reasons for early change of first HAART in HIV-1-infected patients in 7 sites throughout the Caribbean and Latin America. PLoS One. 2010;5(6):e10490.

85. Fink V, Shepherd B, Cesar C, Krolewiecki A, Wehbe F, Cortes C, et al. Cancer in HIV-infected Persons from the Caribbean, Central and South America. submitted.

86. Crabtree-Ramirez B, Caro-Vega Y, Shepherd B, Cesar C, Wehbe F, Cortes C, et al. Prevalence and risk factors associated with late HAART initiation in Latin America and the Caribbean: Late Testers and Late Presenters. submitted. 2010;

87. Kohn LT, Corrigan J, Donaldson MS, America IOM(COQOHCI. To err is human: building a safer health system. National Academies Press; 2000.

88. AIDS Clinical Trials Group [Internet]. [cited 2011 Jan 13];Available from: https://actgnetwork.org/

89. Ottevanger PB, Therasse P, van de Velde C, Bernier J, van Krieken H, Grol R, et al. Quality assurance in clinical trials. Critical Reviews in Oncology/Hematology. 2003 Sep;47(3):213-235.

90. Winchell T. The clinical investigator site audit - process-driven good clinical practice. The Quality Assurance Journal. 2007;11(2):138-142.

91. Li H, Hawlk S, Hanna K, Klein G, Petteway S. Developing and implementing a comprehensive clinical QA audit program. Qual. Assur. J. 2007 6;11(2):128-137.

92. de Lusignan S, van Weel C. The use of routinely collected computer data for research in primary care: opportunities and challenges. Family Practice. 2006 Apr;23(2):253 -263.

93. International Epidemiologic Databases to Evaluate AIDS (IEDEA) [Internet]. [cited 2011 Jan 14];Available from: http://www.iedea-hiv.org/

94. Braitstein P, Brinkhof MWG, Dabis F, Schechter M, Boulle A, Miotti P, et al. Mortality of HIV-1-infected patients in the first year of antiretroviral therapy: comparison between low-income and high-income countries. Lancet. 2006 Mar 11;367(9513):817-824.

95. Egger M, May M, Chêne G, Phillips AN, Ledergerber B, Dabis F, et al. Prognosis of HIV-1-infected patients starting highly active antiretroviral therapy: a collaborative analysis of prospective studies. Lancet. 2002 Jul 13;360(9327):119-129.

96. Zhou J, Kumarasamy N, Ditangco R, Kamarulzaman A, Lee CKC, Li PCK, et al. The TREAT Asia HIV Observational Database: baseline and retrospective data. J. Acquir. Immune Defic. Syndr. 2005 Feb 1;38(2):174-179.

97. Woolley P, Seiler W, Kühn A. Quality control of documents under the constraint of limited resources - maximising the value of QC. Qual. Assur. J. 2004 12;8(4):239-246.

98. TCHARI. Trans-Caribbean HIV/AIDS Research Initiative [Internet]. 2008 May 27 [cited 2010 Nov 5];Available from: http://www.tchari.org/

99. Wolff M, Cortes C, Shepherd B, Beltran C. Long-Term Outcomes of a National Expanded Access Program to Antiretroviral Therapy: The Chilean AIDS Cohort. J Acquir Immune Defic Syndr [Internet]. 2010 Jul 28;Available from: PM:20683194

100. Shepherd BE, Yu C. Accounting for Data Errors Discovered from an Audit in Multiple Linear Regression. Biometrics [Internet]. 2011 Jan 31 [cited 2011 Mar 10];Available from: http://www.ncbi.nlm.nih.gov/pubmed/21281274

101. Favalli G, Vermorken JB, Vantongelen K, Renard J, Van Oosterom AT, Pecorelli S. Quality control in multicentric clinical trials. An experience of the EORTC Gynecological Cancer Cooperative Group. European Journal of Cancer. 2000 Jun;36(9):1125-1133.

102. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)--A metadata-driven methodology and workflow process for providing translational research informatics support. Journal of Biomedical Informatics. 2009 Apr;42(2):377-381.

103. Wagner MM, Hogan WR. The accuracy of medication data in an outpatient electronic medical record. J Am Med Inform Assoc. 1996;3(3):234-244.

104. Hanscom B, Lurie JD, Homa K, Weinstein JN. Computerized questionnaires and the quality of survey data. Spine. 2002 Aug 15;27(16):1797-801.

105. Ryan JM, Corry JR, Attewell R, Smithson MJ. A comparison of an electronic version of the SF-36 General Health Questionnaire to the standard paper version. Qual Life Res. 2002 Feb;11(1):19-26.

106. Hirayama K, Fukuda N, Satoh H, Itoh K, Chiba K, Nakae Y, et al. Checklist for GCP compliance investigation (Medical institution). The Quality Assurance Journal. 2005;9(2):120-139.

107. Dana-Farber/Harvard Cancer Center. Clinical Trials Audit Manual [Internet]. 2009 [cited 2011 Mar 11]; Available from: http://www.dfhcc.harvard.edu/fileadmin/DFHCC_Admin/Clinical_Trials/QACT/Policies_and_Procedures/Audit_Manual_July_2009.doc

108. Bierstaker JL, Burnaby P, Thibodeau J. The impact of information technology on the audit process: an assessment of the state of the art and implications for the future. Managerial Auditing Journal. 2001;16(3):159-164.

109. Abdolmohammadi M, Usoff C. A longitudinal study of applicable decision aids for detailed tasks in a financial audit. International Journal of Intelligent Systems in Accounting, Finance & Management. 2001;10(3):139-154.

110. Lanza R. Take my manual audit, please. Journal of Accountancy [Internet]. 1998 Jun 1;Available from: http://www.thefreelibrary.com/Take+my+manual+audit,+please.-a020854294

111. Glover S, Romeny M. Out in front - four internal audit shops cite technology enhancements. Internal Auditor [Internet]. 1998 Feb [cited 2011 Feb 8];55(1). Available from: http://findarticles.com/p/articles/mi_m4153/is_n1_v55/ai_20568158/

112. ACL Audit Analytics and Continuous Monitoring Software Solutions [Internet]. 2008 [cited 2008 Jul 29];Available from: http://www.acl.com/

113. Audimation Services, Inc. IDEA Data Analysis Software [Internet]. 2007 [cited 2008 Jul 29];Available from: http://www.audimation.com/product_feat_benefits.cfm

114. Financial Accounting Standards Board. FASB Facts About FASB [Internet]. [cited 2011 Feb 7];Available from: http://www.fasb.org/jsp/FASB/Page/SectionPage&cid=1176154526495

115. American Institute of Certified Public Accountants. Audit and Attest Standards [Internet]. 2010 Jun 1 [cited 2011 Feb 7];Available from: http://www.aicpa.org/RESEARCH/STANDARDS/AUDITATTEST/Pages/audit%20and%20attest%20standards.aspx

116. Reinvent Data Limited. TopCAATs - Audit Reinvented [Internet]. [cited 2011 Feb 9];Available from: http://www.topcaats.com/

117. Picalo: Data Analysis and Fraud Detection Toolkit [Internet]. [cited 2008 Jul 29];Available from: http://www.picalo.org/

118. Laubrass. UMT Audit - Industries / Healthcare | Healthcare compliance made easier by using handheld computers for your medical audits. Software developed by experts in the healthcare industry. [Internet]. [cited 2011 Feb 7]; Available from: http://www.laubrass.com/products/?id=19127066

119. Data Quality Assessment Data Verification Templates — UNC Carolina Population Center [Internet]. [cited 2011 Feb 18];Available from: http://www.cpc.unc.edu/measure/tools/monitoring-evaluation-systems/data-quality-assurance-tools/data-quality-assessment-data-verification-templates

120. The United States President's Emergency Plan for AIDS Relief. Data Quality Assurance Tool for Program-Level Indicators [Internet]. Available from: http://www.pepfar.gov/documents/organization/79628.pdf

121. Shakib S, Hoadley P. Auditmaker: a generic tool for clinical audit [Internet]. [cited 2011 Feb 8];Available from: http://www.auditmaker.org/

122. Shakib S, Phillips PA. The Australian Centre for Evidence-based Clinical Practice generic audit tool: Auditmaker for health professionals. J Eval Clin Pract. 2003 May;9(2):259-263.

123. RACGP IT / PEN Computing. Clinical Audit Tool [Internet]. [cited 2011 Feb 7];Available from: http://www.clinicalaudit.com.au/

124. Comparison of Generalized Audit Software [Internet]. [cited 2011 Feb 8];Available from: http://www.auditsoftware.net/documents/GeneralizedAuditSoftware.pdf

125. Duda S, Wehbe F, Gadd C. Desiderata for a computer-assisted audit tool for clinical data source verification audits. Stud Health Technol Inform. 2010;160(Pt 2):894-898.

126. Apache Software Foundation. The Apache HTTP Server Project [Internet]. [cited 2011 Mar 14];Available from: http://httpd.apache.org/

127. PHP: Hypertext Preprocessor [Internet]. [cited 2011 Mar 14]; Available from: http://www.php.net/

128. jQuery Project. jQuery: The Write Less, Do More, JavaScript Library [Internet]. [cited 2011 Mar 14];Available from: http://jquery.com/

129. MySQL. MySQL Database 5.5 [Internet]. [cited 2011 Mar 14];Available from: http://www.mysql.com/products/enterprise/database/

130. Riordan J, Mockler D. Clinical audit in mental health: towards a multidisciplinary approach. John Wiley and Sons; 1997.

131. U.S. Department of Defense. APP AP2.16 Error Classification Codes [Internet]. [cited 2011 Jan 27];Available from: http://www.dla.mil/j-6/dlmso/elibrary/manuals/milstrap/html/040_AP2.16_ERROR_CLASSIFICATION_CODES.htm

132. National Coordinating Council for Medication Error Reporting and Prevention. NCC MERP -- Taxonomy of Medication Errors [Internet]. 1998 [cited 2011 Mar 12];Available from: http://www.nccmerp.org/pdf/taxo2001-07-31.pdf

133. Dovey SM, Meyers DS, Phillips RL, Green LA, Fryer GE, Galliher JM, et al. A preliminary taxonomy of medical errors in family practice. Qual Saf Health Care. 2002 Sep 1;11(3):233-238.

134. Brixey J, Johnson TR, Zhang J. Evaluating a medical error taxonomy. Proc AMIA Symp. 2002;:71-75.

135. Makeham MAB, Dovey SM, County M, Kidd MR. An International Taxonomy For Errors in General Practice: a Pilot Study. The Medical Journal of Australia. 2002 Jul 15;177(2):68-72.

136. Chang A, Schyve PM, Croteau RJ, O'Leary DS, Loeb JM. The JCAHO patient safety event taxonomy: a standardized terminology and classification schema for near misses and adverse events. International Journal for Quality in Health Care. 2005 Apr 1;17(2):95 -105.

137. Kim W, Choi B, Hong E, Kim S, Lee D. A Taxonomy of Dirty Data. Data Mining and Knowledge Discovery. 2003 Jan 1;7(1):81-99.

138. Greenfield L. An (Informal) Taxonomy of Data Warehouse Data Errors [Internet]. 1997 [cited 2008 Apr 3];Available from: http://www.dwinfocenter.org/errors.html

139. McGilvray D. Executing data quality projects: ten steps to quality data and trusted information. Morgan Kaufmann; 2008.

140. Wang R, Storey V, Firth C. A framework for analysis of data quality research. Knowledge and Data Engineering, IEEE Transactions on. 1995;7(4):623-640.

141. Strong DM, Lee YW, Wang RY. Data quality in context. Communications of the ACM. 1997 May;40:103–110.

142. Knight S, Burn J. Developing a framework for assessing information quality on the World Wide Web. Informing Science. 2005;8:159-172.

143. World Health Organization. WHO | WHO case definitions of HIV for surveillance and revised clinical staging and immunological classification of HIV-related disease in adults and children [Internet]. [cited 2011 Mar 14];Available from: http://www.who.int/hiv/pub/vct/hivstaging/en/index.html

144. U.S. Centers for Disease Control and Prevention. 1993 Revised Classification System for HIV Infection and Expanded Surveillance Case Definition for AIDS Among Adolescents and Adults [Internet]. [cited 2011 Mar 14];Available from: http://www.cdc.gov/mmwr/7preview/mmwrhtml/00018871.htm

145. Komaroff A. The variability and inaccuracy of medical data. Proceedings of the IEEE. 1979;67(9):1196-1207.

146. Rethans JJ, Martin E, Metsemakers J. To what extent do clinical notes by general practitioners reflect actual medical performance? A study using simulated patients. Br J Gen Pract. 1994 Apr;44(381):153-156.

147. Luck J, Peabody JW, Dresselhaus TR, Lee M, Glassman P. How well does chart abstraction measure quality? A prospective comparison of standardized patients with the medical record. The American Journal of Medicine. 2000 Jun;108(8):642-649.

148. Chengalur-Smith IN, Ballou DP, Pazer HL. The Impact of Data Quality Information on Decision Making: An Exploratory Analysis. IEEE Trans. on Knowl. and Data Eng. 1999 Nov;11:853–864.